**Improving Adaptive Argumentation Learning through Artificial Intelligence**

D I S S E R T A T I O N

of the University of St.Gallen,

School of Management,

Economics, Law, Social Sciences,

International Affairs and Computer Science,

to obtain the title of

Doctor of Philosophy in Management

submitted by

**Thiemo Wambsganss**

from

from Germany

Approved on the application of

**Prof. Dr. Jan Marco Leimeister**

and

**Prof. Dr. Matthias Söllner**

Dissertation no. 5259

Difo-Druck GmbH, Untersiemau 2022

The University of St.Gallen, School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science, hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St.Gallen, May 24, 2022

The President:

Prof. Dr. Bernhard Ehrenzeller

*To my family*

# Acknowledgements

This dissertation is the result of three and a half years' work under the supervision of Prof. Dr. Jan Marco Leimeister at the Institute of Information Management at the University of St. Gallen. Hence, I would like to thank Prof. Dr. Jan Marco Leimeister for his supervision of my thesis, the reflection, and broader perspectives on my work throughout the three and a half years of my doctorate journey. Furthermore, I would like to thank Prof. Dr. Matthias Söllner for his feedback, supervision, and mentoring during my doctorate. He not only supported my research with ongoing ideas and resources, but also helped me to grow as a researcher and opened up several opportunities.

Also, I would like to thank Prof. Ken Koedinger for hosting me one year at his lab at Carnegie Mellon University. The collaboration not only contributed significantly to my scientific learning curve but also helped me a lot to sharpen my work and to grow personally. Next, I would like to thank Prof. Dr. Hansjörg Fromm for his deep interest in and our collaborations on Natural Language Processing. The mentoring from Hansjörg, especially before and at the beginning of my doctorate, provided me with a solid foundation for my research endeavor.

Furthermore, I am particularly grateful to my co-authors and my colleagues at the Institute of Information Management and the University of Kassel. Especially, I would like to thank Christian Engel and Christian Marheine with whom I started and successfully accomplished this doctorate journey. Special thanks go to Dr. Rainer Winkler who not only brought a nice positive atmosphere into my daily work but from who I also learned a lot about pedagogical scenarios and technology-mediated learning. Moreover, I would like to thank Matthias Cetto and Anuschka Schmitt for our fruitful collaborations, our innovative discussions, and our common research interest.

I express my special thanks to my mother Helga Kühner-Wambsganß and my father Wolfgang Wambsganß, who always supported me during my studies and my doctorate journey. Moreover, I would like to thank my sister Dr. Janna Wambsganß for her support and academic advice on my work from a natural science perspective. Also, I would like to thank Dr. Mujtaba Syed for his feedback and guidance.

Finally, I am very grateful to my girlfriend Jin-ah Lee, who not only ensured that I kept a healthy work-life balance during my doctorate, but also for her tremendous support, advice, and discussions on personal and job-related challenges.

Pittsburgh, December 2021

Thiemo Wambsganss

# List of Publications

During my time as a research assistant and Ph.D. student at the Institute of Information Management (IWI) at the University of St. Gallen (HSG) and as a visiting research collaborator at the Human-Computer Interaction Institute (HCII) at the Carnegie Mellon University, I authored and co-authored the following publications under the supervision of Prof. Dr. Jan Marco Leimeister (PhD supervisor). The majority of the articles are related to my dissertation topic, and hence overlap with the contents of this thesis. I provide brief notes in each relevant area of my dissertation emphasizing publications that have had a major impact on the content of that section. Following is a chronological list of all publications, beginning with the most recent (under review, 2022, 2021, 2020, 2019):

# Under Review

Thiemo Wambsganss, Matthias Söllner, Andreas Janson, Ken Koedinger, and Jan Marco Leimeister. 2022b. Improving Argumentation Based on a Dynamic Feedback System Created with Machine Learning. *Under review at Information Systems Research*, pages 1–38

Thiemo Wambsganss. 2022. Nudging Towards Self-Regulated Argumentation Learning: Designing Intelligent Argumentation Writing Support Systems. *Institute of Information Management at the University of St.Gallen Working Paper, St. Gallen, Switzerland*, pages 1–30

# Publications

## 2022

Thiemo Wambsganss, Matthias Soellner, Kenneth R Koedinger, and Jan Marco Leimeister. 2022a. Adaptive empathy learning support in peer review scenarios. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–

8760, Dublin, Ireland. Association for Computational Linguistics

## 2021

Thiemo Wambsganss, Christian Engel, and Hansjörg Fromm. 2021a. Improving Explainability and Accuracy through Feature Engineering : A Taxonomy of Features in NLP-based Machine Learning. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas

Thiemo Wambsganss, Anuschka Schmitt, Thomas Mahnig, Anja Ott, Sigita Söllner, Ngoc Anh Ngo, Jerome Geyer-klingeberg, Janina Nakladal, and Jan Marco Leimeister. 2021e. The Potential of Technology-Mediated Learning Processes : A Taxonomy and Research Agenda for Educational Process Mining. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas

Anuschka Schmitt, Thiemo Wambsganss, Andreas Janson, and Matthias Söllner. 2021. Towards a Trust Reliance Paradox ? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas

Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Lernaing : A Taxonomy of Conversational Agents in Education. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas

Thiemo Wambsganss, Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. 2021b. AI-based Argumentation Tutoring – A Novel System Class to Improve Learners' Argumentation Skills. In *Annual Meeting of the Academy of Management (AOM). - Virtual Conference, 2021*, volume 2021, page 10233

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021d. Supporting Cognitive and Emotional Empathic Writing of Students. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4063–4077

Thiemo Wambsganß, Leonie Haas, and Matthias Söllner. 2021. Towards the Design of a Student-Centered Question-Answering System in Educational Settings. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*

Thiemo Wambsganss and Christian Engel. 2021. Using Deep Learning for Extracting User-Generated Knowledge from Web Communities. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*, pages 1–16

Arne Gruettner, Thiemo Wambsganss, and Andrea Back. 2021. From data to dollar: Using the wisdom of an online tipster community to improve sports betting returns. *European Journal of International Management*, 15(2-3):314–338

Thiemo Wambsganss. 2021. Designing Adaptive Argumentation Learning Systems Based on Artificial Intelligence. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*

Thiemo Wambsganss, Tobias Kueng, Matthias Söllner, and Jan Marco Leimeister. 2021c. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA

Thiemo Wambsganss, Anne Höch, Naim Zierau, and Matthias Söllner. 2021a. Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design. In *Proceedings of the 16th International Conference on Wirtschaftsinformatik*, Essen, Germany

Thiemo Wambsganss, Sebastian Guggisberg, and Matthias Söllner. 2021b. ArgueBot: A Conversational Agent for Adaptive Argumentation Feedback. In *16th International Conference on Wirtschaftsinformatik*, Essen, Germany

Thiemo Wambsganß, Florian Weber, and Matthias Söllner. 2021. Designing an Adaptive Empathy Learning Tool. In *16th International Conference on Wirtschaftsinformatik*, Essen, Germany

Severin Landolt, Thiemo Wambsganß, and Matthias Söllner. 2021. A taxonomy for deep learning in natural language processing. In *Proceedings of the Annual Hawaii International Conference on System Sciences*

Thiemo Wambsganss, Florian Weber, and Matthias Söllner. 2021f. Design and Evaluation of an Adaptive Empathy Learning Tool. In *Hawaii International Conference on System Sciences*

## 2020

Thiemo Wambsganss, Matthias Söllner, and Jan Marco Leimeister. 2020d. Design and Evaluation of an Adaptive Dialog-Based Tutoring System for Argumentation Skills. In *International Conference on Information Systems (ICIS)*, Hyderabad, India

Naim Zierau, Thiemo Wambsganss, Andreas Janson, Sofia Schöbel, and Jan Marco Leimeister. 2020. The Anatomy of User Experience with Conversational Agents: A Taxonomy and Propositions of Service Clues. In: International Conference on Information Systems (ICIS). Anatomy of User Experience with Conversational Agents The Anatomy of User Experience with. In *International Conference on Information Systems (ICIS)*, Hyderabad, India

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020c. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*, Barcelona, Spain

Thiemo Wambsganss, Rainer Winkler, Pascale Schmid, and Matthias Söllner. 2020g. Unleashing the Potential of Conversational Agents for Course Evaluations: Empirical Insights from a Comparison with Web Surveys. In *Twenty-Eighth European Conference on Information Systems (ECIS2020)*, pages 1–18, Marrakesh, Morocco

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020b. AL : An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA. Association for Computing Machinery

Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020h. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*

Thiemo Wambsganss, Florian Weber, Tobias Staufenberg, Leopold Bott, and Matthias Söllner. 2020e. Systematische Entwicklung eines Lerntools zur Erhöhung der Argumentationsfähigkeiten von Studierenden. In *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany

Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020a. Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach. In *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany

Thiemo Wambsganss, Rainer Winkler, Pascale Schmid, and Matthias Söllner. 2020f. Designing a Conversational Agent as a Formative Course Evaluation Tool. In *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany

Arne Gruettner, Min Vitisvorakarn, Thiemo Wambsganss, Roman Rietsche, and Andrea Back. 2020. The new window to athletes' soul - What social media tells us about athletes' performances. *Proceedings of the Annual Hawaii International Conference on System Sciences*

## 2019

Thiemo Wambsganss and Roman Rietsche. 2019. Towards designing an adaptive argumentation learning tool. In *40th International Conference on Information Systems, ICIS 2019*, Munich, Germany

Hansjörg Fromm, Thiemo Wambsganss, and Matthias Söllner. 2019. Towards a Taxonomy of Text Mining Features. In *In Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm, Sweden

# Contents

**Table of Contents** xiii

**List of Figures** xix

**List of Tables** xxiii

**List of Abbreviations** xxv

**abstract** xxv

**zusammenfassung** xxvi

**1 Introduction** 1

1.1 Problem Statement . . . 1
1.2 Solution Statement and Research Questions . . . 4
1.3 Structure of the Dissertation . . . 12
1.4 Overview of Publications in Relation to the Research Questions . . . 13

**2 Theoretical Background** 17

2.1 Technology-Mediated Learning . . . 17
2.2 Learning Context: Argumentation Skill Learning . . . 20
2.2.1 Learning Goals and Learning Outcomes . . . 20
2.2.2 The Concept of Argumentation . . . 21
2.2.3 Argumentation Theory . . . 22
2.2.4 Argumentation Models . . . 23
2.2.5 Argumentation Skill Learning . . . 24
2.2.6 Constructivist Learning Context . . . 25
2.3 Learning Method Structures: Technology-Mediated Argumentation Skill Learning . . . 29
2.3.1 Computer-Supported Collaborative Learning . . . 30
2.3.2 Intelligent Tutoring Systems . . . 30
2.3.3 Pedagogical Conversational Agents . . . 32
2.3.4 Argumentation Learning Systems . . . 33
2.4 Learning Process: Learning to Argue vs Arguing to Learn . . . 49
2.5 Technical Background . . . 50
2.5.1 Natural Language Processing . . . 50
2.5.2 Machine Learning . . . 53
2.5.3 Supervised Classification Approach . . . 54
2.5.4 Corpus Design . . . 56
2.5.5 Evaluation Metrics . . . 58

2.5.6 Argumentation Mining . . . . . 59

**3 Methodological Background 61**
3.1 Underlying Research Epistemology . . . . . 61
3.2 Literature Review . . . . . 63
3.3 Taxonomy Development . . . . . 64
3.4 Quantitative Methods . . . . . 66
3.5 Qualitative Methods . . . . . 68
3.6 Design Science Research . . . . . 70
3.6.1 Theory in Design Science Research . . . . . 70
3.6.2 Process Models to Guide Design Science Research . . . . . 73
3.6.3 Cross-Industry Standard Process for Data Mining . . . . . 77

**4 Developing an Understanding of Natural Language Processing for ML-based Information Systems 79**
4.1 Taxonomy of Features for NLP-based Machine Learning . . . . . 80
4.1.1 Introduction . . . . . 80
4.1.2 Theoretical Background . . . . . 82
4.1.3 Methodology . . . . . 85
4.1.4 Taxonomy of NLP Features for ML-based IS . . . . . 89
4.1.5 Cluster Heatmapping of NLP Features Across Studies . . . . . 94
4.1.6 Discussion . . . . . 96
4.1.7 Conclusion . . . . . 98
4.2 Utilizing Transfer Learning for Argumentation Identification . . . . . 100
4.2.1 Introduction . . . . . 100
4.2.2 Theoretical Background . . . . . 101
4.2.3 Methodology . . . . . 104
4.2.4 Implementation . . . . . 105
4.2.5 Discussion . . . . . 110
4.2.6 Conclusion . . . . . 111

**5 Designing AI-based Argumentation Writing Support Systems 112**
5.1 Requirements for the Design of AI-based Argumentation Writing Support Systems . . . . . 116
5.1.1 Literature Requirements for AI-based Argumentation Writing Support Systems . . . . . 116
5.1.2 User Requirements for AI-based Argumentation Writing Support Systems . . . . . 118
5.1.3 Deriving Design Principles for AI-based Argumentation Writing Support Systems . . . . . 119

5.1.4 Conclusion . . . 125
5.2 Design and Implementation of an AI-based Argumentation Writing Support System . . . 126
5.2.1 Proposed Architecture of AI-based Argumentation Writing Support Systems . . . 126
5.2.2 Annotating and Developing a Corpus of Argumentation Annotated Student-Written Peer Reviews . . . 128
5.2.3 Instantiation of the AI-based Argumentation Writing Support System ArgueLearn . . . 140
5.2.4 Argumentation Mining Algorithm of ArgueLearn . . . 143
5.3 Demonstration and Evaluation of an AI-based Argumentation Writing Support System . . . 148
5.3.1 Pre-Study: Design Evaluation . . . 148
5.3.2 Study 1: Proof of Concept Evaluation in a Laboratory Experiment 150
5.3.3 Study 2: Proof of Value Evaluation of ArgueLearn in a Field Experiment . . . 159
5.4 Discussion . . . 165
5.4.1 Theoretical and Practical Contributions . . . 166
5.4.2 Limitations and Future Work . . . 167
5.5 Conclusion . . . 167

**6 Designing Conversational Argumentation Learning Systems 169**
6.1 Introduction . . . 171
6.2 Theoretical Background . . . 174
6.3 Requirements for the Design of Conversational Argumentation Learning Systems . . . 175
6.3.1 Deriving Literature Requirements for Conversational Argumentation Learning Systems . . . 175
6.3.2 Deriving User Requirements for Conversational Argumentation Learning Systems . . . 176
6.3.3 Deriving Design Principles for Conversational Argumentation Learning Systems . . . 176
6.4 Design and Implementation of a Conversational Argumentation Learning System . . . 177
6.4.1 Instantiation of the Conversational Argumentation Learning System ArgueTutor . . . 179
6.4.2 Argumentation Mining Algorithm of ArgueTutor . . . 183
6.5 Demonstration and Evaluation of a Conversational Argumentation Learning System . . . 184

6.5.1 Pre-Study . . . 185
6.5.2 Study 1: Proof of Concept Evaluation of ArgueTutor . . . 186
6.5.3 Study 2: Experimental Evaluation of Adaptive Conversational Feedback . . . 194
6.6 Discussion . . . 197
6.6.1 Theoretical and Practical Contributions . . . 198
6.6.2 Limitations and Future Work . . . 198
6.7 Conclusion . . . 199

**7 Nudging Students Towards Self-Regulated Argumentation Learning 200**
7.1 Introduction . . . 202
7.2 Theoretical Background . . . 204
7.3 Design and Development of ArgumentFeedback . . . 206
7.3.1 Deriving Requirements for Nudging Students Towards more Persuasive Writing . . . 206
7.3.2 User Interaction of ArgumentFeedback . . . 209
7.3.3 Annotating and Developing a Corpus of Argumentation-Annotated Student-Written Business Pitches . . . 211
7.3.4 Modeling Argumentation Structures in Persuasive Business Pitches 225
7.4 Demonstration and Evaluation . . . 227
7.4.1 Study 1: Qualitative Design Evaluation . . . 227
7.4.2 Study 2: Proof of Concept Laboratory Experiment . . . 229
7.4.3 Study 3: Proof of Value Field Experiment . . . 231
7.5 Discussion . . . 240
7.5.1 Theoretical and Practical Contributions . . . 241
7.5.2 Limitations and Future Work . . . 243
7.6 Conclusion . . . 244

**8 Summary of Contributions and Areas for Future Research 245**
8.1 Theoretical Contributions of the Dissertation . . . 245
8.1.1 Design of NLP-based Adaptive Educational Learning Tools . . . 245
8.1.2 Design of Adaptive Argumentation Learning Systems . . . 247
8.1.3 Components of a Design Theory for AI-based Argumentation Writing Support Systems . . . 250
8.1.4 Components of a Design Theory for Conversational Argumentation Learning Systems . . . 254
8.1.5 Nudging for Self-Regulated Argumentation Learning . . . 259
8.2 Practical Contributions of the Dissertation . . . 261
8.2.1 Software Instantiations . . . 261
8.2.2 Annotation Schemes and Argumentation Annotated Corpora . . . 261

8.2.3 Findings for Educators . . . 263
8.3 Limitations and Directions for Future Research . . . 263
8.3.1 Research on Novel Pedagogical Scenarios for Argumentation Learning Systems . . . 265
8.3.2 Research on the Learner-Centered Design of Argumentation Learning Systems . . . 268
8.3.3 Research on Recent Technological Advances for Argumentation Learning Systems . . . 269
8.3.4 Research on Ethics for AI-based Learning Tools . . . 270

**References** **272**

**A Appendix** **305**
A.1 Guidelines for the Annotation of Argumentation Structures in Student-Written Peer Reviews . . . 305
A.1.1 Introduction . . . 305
A.1.2 Argument components . . . 306
A.1.3 Limits of argumentation components . . . 306
A.1.4 Annotation of opinions and premises in reviews . . . 313
A.1.5 Argumentative relations . . . 313
A.2 Guideline for the Annotation of Argument Structure and Quality in Business Pitches . . . 314
A.2.1 Introduction . . . 314
A.2.2 Argument components . . . 317
A.2.3 Relationships of argumentation components . . . 319
A.2.4 Rules for annotation of argumentation components . . . 320
A.2.5 Quality of argumentation components . . . 324

# List of Figures


Fig. 1: Focus of the dissertation based on the conceptual model of technology-mediated learning by Gupta and Bostrom (2009).
Source: Gupta and Bostrom (2009) . . . . . 5

Fig. 2: Overview of the dissertation structure in eight consecutive sections.
Source: Own illustration . . . . . 12

Fig. 3: Focus of the dissertation based on the technology-mediated learning framework by Gupta and Bostrom (2009).
Source: Gupta and Bostrom (2009) . . . . . 19

Fig. 4: Aristotle's theory of persuasion based on logos, ethos, and pathos.
Source: Own illustration . . . . . 22

Fig. 5: The six elements an argument can be broken down according to the Toulmin model.
Source: Toulmin (2003) . . . . . 23

Fig. 6: Basic argumentation model with the logos components claim and premise.
Source: Own illustration . . . . . 24

Fig. 7: Basic concept of the representational guidance approach of Suthers (2003); different representations of argumentation structures are manipulated to guide the argumentation skill learning of students.
Source: Scheuer (2015); Suthers (2003) . . . . . 35

Fig. 8: Screenshot of the Belvedere system (Suthers et al. 2001). It uses a node-and-link graph and a table to express hypotheses-evidence relationships.
Source: Suthers (2003); Scheuer (2015) . . . . . 36

Fig. 9: Learning tool by Weinberger et al. (2010) to support the creation of argumentative messages through a form-like interface in a forum discussion.
Source: Weinberger et al. (2010); Scheuer (2015) . . . . . 38

Fig. 10: AcademicTalk system by McAlister et al. (2004): Students are provided with sentence openers to facilitate a discussion.
Source: McAlister et al. (2004); Scheuer (2015) . . . . . 38

Fig. 11: LASAD user interface adapted from Scheuer et al. (2012a): The system combines graph-based representations with sentence openers based on a domain-independent ontology.
Source: Scheuer et al. (2012a) . . . . . 40

Fig. 12: LASAD user interface adapted from Scheuer et al. (2012a): The system provides users with feedback messages based on detected patterns.
Source: Scheuer et al. (2012a) . . . . . 41

Fig. 13: Screenshot of the argumentation mining tool Margot.
Source: Lippi and Torroni (2016b) . . . . . 41

Fig. 14: Interface of ArguminSci for argument claim detection in scientific writing.
Source: Lauscher et al. (2019) . . . . . 42
Fig. 15: Argument Targer based on different argumentation mining models for claim and premise highlighted.
Source: Chernodub et al. (2019) . . . . . 42
Fig. 16: Overview of NLP techniques as standard text processing for ML-based systems.
Source: Own illustration . . . . . 51
Fig. 17: Overview of the levels of linguistics illustrated with examples.
Source: Own illustration . . . . . 52
Fig. 18: Structure of available data for supervised classification.
Source: Bird et al. (2009) . . . . . 54
Fig. 19: Overview of supervised classification framework with training and prediction.
Source: Bird et al. (2009) . . . . . 55
Fig. 20: Iterative taxonomy development process.
Source: Nickerson et al. (2013) . . . . . 65
Fig. 21: Classification of design science knowledge contribution types by the solution maturity and the application domain maturity.
Source: Gregor and Hevner (2013) . . . . . 72
Fig. 22: Process perspective of desgin science research according to Peffers et al. (2007).
Source: Peffers et al. (2007) . . . . . 75
Fig. 23: Overview of the three-cycle view to guide the iterative process of design science research.
Source: Hevner (2007) . . . . . 75
Fig. 24: Structure of a data mining research iteration according to the CRISP-DM Model.
Source: Chapman et al. (2000) . . . . . 78
Fig. 25: Illustration of pre-processing, feature engineering and NLP in an intelligent IS.
Source: Own illustration . . . . . 82
Fig. 26: Overview on analyzed NLP features across domains.
Source: Own illustration . . . . . 90
Fig. 27: Taxonomy of NLP features in an NLP feature canvas.
Source: Own illustration . . . . . 90

Fig. 28: Overview on the sample of 211 NLP research, the frequency distribution of text features for the linguistic analysis level, granularity level, and information source.
Source: Own illustration . . . . . 95
Fig. 29: Illustration of the unclustered and clustered data matrix in a heat map. Lines of the matrix represent publications, columns represent features.
Source: Own illustration . . . . . 95
Fig. 30: Outline of the proposed learning architecture for argumentation identification.
Source: Adopted from Devlin et al. (2018) . . . . . 108
Fig. 31: Design science research methodology for the design of AI-based argumentation writing support systems.
Source: Own illustration based on Peffers et al. (2007) . . . . . 114
Fig. 32: Overview of the proposed architecture of an AI-based argumentation writing support system.
Source: Own illustration . . . . . 127
Fig. 33: Overview on the process of collecting a corpus of 1,000 annotated peer reviews written by students.
Source: Own illustration . . . . . 132
Fig. 34: Argumentation annotation scheme including argument components (claim, premise) and argumentative relations (support or attack).
Source: Own illustration . . . . . 133
Fig. 35: Examples of argumentation relations that could be found in the corpus. The arrow represents a relationship, which might be support or attack. The rectangle signifies claim or premise, which is an argument component.
Source: Own illustration . . . . . 134
Fig. 36: Fully annotated example of a peer review according to the annotation scheme and guidelines. The left column represents the claims (C1 - C6), while the premises are listed on the right (P1 - P7). N signifies a non-argumentative text unit.
Source: Own illustration . . . . . 136
Fig. 37: Early design concept of ArgueLean for receiving initial user feedback in the pre-study.
Source: Own illustration . . . . . 140
Fig. 38: Screenshot of the second version of ArgueLearn's interface with instantiated functionalities (F1 to F7). This version of ArgueLearn was used for evaluating the design principles in a laboratory experiment (study 1).
Source: Own illustration . . . . . 141

Fig. 39: Screenshot of the final version of ArgueLearn's interface with instantiated functionalities (F1 to F6). This version was instantiated and evaluated in a field experiment (study 2).
Source: Own illustration . . . . . 142
Fig. 40: Top: screenshot of detailed discourse feedback of unsupported claim. Below: screenshot of exemplary explanations and details of ArgueLearn.
Source: Own illustration . . . . . 143
Fig. 41: Overview of experimental procedure of study 1 to evaluate if an AI-based argumentation writing support system based on my design principles improves persuasive writing more so than traditional argumentation scripting.
Source: Own illustration . . . . . 150
Fig. 42: Screenshot of discussion scripting approach: a user enters a claim and the corresponding premise into the input masks and adds it to the final text.
Source: Own illustration . . . . . 152
Fig. 43: Results on formal (left) and perceived (right) quality of argumentation between both tools.
Source: Own illustration . . . . . 156
Fig. 44: Overview of experimental procedure of study 2 to evaluate if an AI-based writing support system based on my principles, in comparison to static argumentation feedback, leads to better learning of argumentation.
Source: Own illustration . . . . . 159
Fig. 45: Screenshot of the task interface: students were asked to write a persuasive peer review.
Source: Own illustration . . . . . 160
Fig. 46: Screenshot of treatment group 1 (TG1): students received general argumentation feedback with hints and recommendations on what to improve based on the model of Toulmin (1984).
Source: Own illustration . . . . . 162
Fig. 47: Screenshot of treatment group 2 (TG2): students received AI-based argumentation writing support by ArgueLearn following my four design principles.
Source: Own illustration . . . . . 163
Fig. 48: Overview of my methodology based on the design science research approach following Hevner (2007).
Source: Own illustration . . . . . 172

Fig. 49: Overview of the derived design principles for conversational argumentation learning systems according to Gregor et al. (2020).
Source: Own illustration . . . . . 178
Fig. 50: Basic user interaction concept of my conversational argumentation learning system ArgueTutor: a student is adaptively tutored through a writing exercise with individual argumentation feedback.
Source: Own illustration . . . . . 178
Fig. 51: Different screenshots of the adaptive conversational argumentation learning system ArgueTutor: a user conducts a certain writing exercise and receives adaptive tutoring and feedback on the argumentation quality of her text.
Source: Own illustration . . . . . 179
Fig. 52: Screenshot of adaptive learning mode (F5) and casual chat mode (F7) of the conversational argumentation learning system ArgueTutor.
Source: Own illustration . . . . . 180
Fig. 53: Screenshot of the interface of ArgueTutor with exemplars of design features in a mobile and desktop version.
Source: Own illustration . . . . . 181
Fig. 54: The dialog flow of ArgueTutor consists of an adaptive argumentation learning mode (mode 1) and a casual chat function (mode 2).
Source: Own illustration . . . . . 182
Fig. 55: Basic user interaction concept of alternative argumentation learning tool: students receive theory input and general recommendations on the argumentation of a given text.
Source: Own illustration . . . . . 188
Fig. 56: Basic concept of ArgumentFeedback: students are conducting a persuasive writing exercise in a common writing editor and are nudged with an intelligent self-evaluation towards more argumentative writing.
Source: Own illustration . . . . . 206
Fig. 57: Screenshot of the AI-based argumentation writing support system ArgumentFeedback used in this study: a student conducts a persuasive writing exercise and is nudged with intelligent argumentation self-evaluation based on the argumentation quality of her text to increase augmentation skill learning.
Source: Own illustration . . . . . 210

Fig. 58: Additional functionalities of ArgumentFeedback. Left: argumentation theory introduction and goals setting based on the Toulmin model Middle: transparent and adaptive explanations of summarizing scores Right: help button for further information and learning support.
Source: Own illustration . . . . . 212
Fig. 59: Overview of the argumentation annotation scheme for business model pitches, including argument components *(major claim, claim, premise)*, argumentative relations *(support)*, and persuasiveness scores *(specificity, evidence, strength, relevance)*.
Source: Own illustration . . . . . 215
Fig. 60: Annotated example of a pitch business model pitch according to the annotation scheme and guidelines.
Source: Own illustration . . . . . 216
Fig. 61: Overview of the annotation scheme including the argumentative components and the persuasiveness scores.
Source: Own illustration . . . . . 217
Fig. 62: Overview of the three experimental groups: participants of the control group (CG) received argumentation self-monitoring based on syntactical rules while conducting an argumentative writing exercise. Students in treatment group 1 (TG1) conducted the same exercise and were nudged with intelligent argumentation self-monitoring based on the trained ML model. Participants in treatment group 2 (TG2) were nudged with an intelligent self-evaluation interface based on social comparison and an adaptive feedback message based on the trained ML model.
Source: Own illustration . . . . . 232
Fig. 63: Results on formal (left) and perceived (right) quality of argumentation between the three groups (CG: argumentation self-monitoring, TG1: intelligent argumentation self-monitoring nudge, TG2: intelligent argumentation self-evaluation nudge).
Source: Own illustration . . . . . 236
Fig. 64: Results on self-reported argumentation of participants (left) and subjective argumentation changes of participants (right) between the three groups (CG: argumentation self-monitoring, TG1: intelligent argumentation self-monitoring nudge, TG2: intelligent argumentation self-evaluation nudge).
Source: Own illustration . . . . . 239

# List of Tables


Tab. 1: Overview of publications related to the dissertation according to the research questions.
Source: Own illustration . . . . . 16

Tab. 2: Revised taxonomy of learning outcomes according to Krathwohl (2002).
Source: Krathwohl (2002) . . . . . 20

Tab. 3: Exemplary learning activities by mode of engagement according to the ICAP Framework from Chi and Wylie (2014).
Source: Chi and Wylie (2014) . . . . . 27

Tab. 4: Overview of different IT-based argumentation learning approaches.
Source: Own illustration . . . . . 48

Tab. 5: NLP preprocessing steps for machine learning-based IS.
Source: Andrews and Fox (2007) . . . . . 52

Tab. 6: Non-exhaustive overview on classifier learning categories.
Source: Michie et al. (1994); Opitz and Maclin (1999) . . . . . 56

Tab. 7: Overview on the eight components of an IS design theory.
Source: Jones and Gregor (2007) . . . . . 71

Tab. 8: The components of design principles to structure generalizable design guidelines.
Source: Gregor et al. (2020) . . . . . 73

Tab. 9: Overview of the four consecutive research steps for deriving a taxonomy for NLP-based features in adaptive IS.
Source: Own illustration . . . . . 86

Tab. 10: Feature examples categorized by linguistic analysis level, granularity level, external sources, dimensionality and representation.
Source: Own illustration . . . . . 92

Tab. 11: A not exhaustive overview of approaches and level of analysis in argumentation mining literature.
Source: Own illustration . . . . . 103

Tab. 12: Results of binary classification for argumentation identification based on current literature.
Source: Own illustration . . . . . 107

Tab. 13: Achieved results for different domain corpora with the same BERT learning model.
Source: Own illustration . . . . . 110

Tab. 14: Overview on hits and relevant for AI-based argumentation learning tools in different databases.
Source: Own illustration . . . . . 117

Tab. 15: Inter-rater agreement of argument component annotations.
Source: Own illustration . . . . . 137
Tab. 16: Inter-rater agreement of argumentative relation annotations.
Source: Own illustration . . . . . 137
Tab. 17: Distribution of *sentences* and *tokens* in the created corpus. Mean, standard deviation, minimum, maximum and median refer to the number of sentences and tokens, respectively, per document.
Source: Own illustration . . . . . 138
Tab. 18: Distribution of types of *argument components* in the created corpus.
Source: Own illustration . . . . . 138
Tab. 19: Distribution of types of *argumentative relations* in the created corpus.
Source: Own illustration . . . . . 138
Tab. 20: Distribution of patterns of *support relationships* in the created corpus. The columns denote the percentage of claims and premises, respectively, that are supported by the specified number of premises.
Source: Own illustration . . . . . 139
Tab. 21: Features used for argument component identification according to Stab and Gurevych (2017) and the NLP feature taxonomy from subsection 4.1.
Source: Stab and Gurevych (2017) . . . . . 144
Tab. 22: Features used for argument relation identification the NLP feature taxonomy from subsection 4.1.
Source: Stab and Gurevych (2017) . . . . . 146
Tab. 23: Overview of demonstration and evaluation of the AI-based argumentation writing support system guided by Venable et al. (2016).
Source: Own illustration . . . . . 149
Tab. 24: Results of the technology acceptance of ArgueLearn and the reference tool on a 1 - 5 Likert Scale (1: high, 5: low).
Source: Own illustration . . . . . 155
Tab. 25: Representative examples of qualitative user responses from users of ArgueLearn after study 1.
Source: Own illustration . . . . . 158
Tab. 26: Mean and standard derivation of formal and perceived argumentation between the three groups of the field experiment (study 2).
Source: Own illustration . . . . . 164
Tab. 27: Mean and standard derivation of formal and perceived argumentation between the three groups of the field experiment (study 2).
Source: Own illustration . . . . . 164

Tab. 28: Results of formal and perceived quality of argumentation between both tools.
Source: Own illustration . . . 191
Tab. 29: Results of the perceived ease of use and level of enjoyment ArgueTutor and the reference tool on a 1 - 5 Likert Scale.
Source: Own illustration . . . 192
Tab. 30: Representative examples of qualitative user responses.
Source: Own illustration . . . 193
Tab. 31: Measured mean and standard derivation on a 1-5 Likert scale (1: high, 5: low); *p <0.05, **p <0.01.
Source: Own illustration . . . 196
Tab. 32: Theory-based design requirements following my findings in section 5 on how to build an intelligent argumentation writing support system.
Source: Own illustration . . . 209
Tab. 33: Summary of the descriptions of the persuasiveness scores on the different argumentative components.
Source: Own illustration . . . 218
Tab. 34: Description of the different *specificity* scores for major claims.
Source: Own illustration . . . 219
Tab. 35: Description of the *evidence* score for major claims and claims.
Source: Own illustration . . . 219
Tab. 36: Description of the *relevance* score for premises.
Source: Own illustration . . . 220
Tab. 37: Description of the *strength* score for premises.
Source: Own illustration . . . 220
Tab. 38: Description of the *strength* score for the pitch.
Source: Own illustration . . . 220
Tab. 39: Inter-annotator agreement of argument component annotations.
Source: Own illustration . . . 222
Tab. 40: Inter-annotator agreement of argumentative relation annotations.
Source: Own illustration . . . 222
Tab. 41: Inter-annotator agreement of the qualitative argumentation labels.
Source: Own illustration . . . 223
Tab. 42: Distribution of *sentences* and *tokens* in the created corpus. Mean, standard deviation, minimum, maximum, and median refer to the number of sentences and tokens, respectively, per document.
Source: Own illustration . . . 224
Tab. 43: Overview on descriptive statistics for the types of *argument components*.
Source: Own illustration . . . 224

Tab. 44: Types of *argumentative relations*.
Source: Own illustration . . . . . 225
Tab. 45: Statistics on the qualitative labels.
Source: Own illustration . . . . . 225
Tab. 46: Distribution of patterns of *support relationships* in the created corpus. The columns denote the percentage of claims and premises, respectively, that are supported by the specified number of premises.
Source: Own illustration . . . . . 226
Tab. 47: Overview of the demonstration and experimental evaluation of ArgumentFeedback in three subsequent studies.
Source: Own illustration . . . . . 228
Tab. 48: Results of formal quality of argumentation between the groups on a 1-7 Likert scale (***p<0.001, **p<0.01, *p<0.05).
Source: Own illustration . . . . . 237
Tab. 49: Results of perceived quality of argumentation between the groups on a 1-7 Likert scale, 1: not very persuasive, 7: very persuasive, 4: neutral (***p<0.001, **p<0.01, *p<0.05).
Source: Own illustration . . . . . 237
Tab. 50: Results of self-reported argumentation between the groups (***p<0.001, **p<0.01 *p <0.05).
Source: Own illustration . . . . . 239
Tab. 51: Results of self-reported argumentative changes between the groups (***p<0.001, **p<0.01 *p<0.05).
Source: Own illustration . . . . . 240
Tab. 52: Representative examples of qualitative user responses.
Source: Own illustration . . . . . 241
Tab. 53: Design theory for AI-based argumentation writing support systems to enable individuals to train their argumentation skills irrespective of instructor, time, and place according to Jones and Gregor (2007).
Source: Own illustration . . . . . 254
Tab. 54: Design theory for AI-based conversational argumentation learning systems according to Jones and Gregor (2007).
Source: Own illustration . . . . . 258

# List of Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| ANOVA | Analysis of Variance |
| AWSS | Argumentation Writing Support System |
| BOW | Bag-of-Words |
| CALS | Conversational Argumentation Learning System |
| DSR | Design Science Research |
| CRISP-DM | Cross Industry Standard Process for Data Mining Models |
| IT | Information Technology |
| ICAP framework | Interactive, Constructive, Active and Passive Framework |
| HCI | Human Computer Interaction |
| LI | Literature Issues |
| LSTM | Long Short-Term Memory |
| BiLSTM-CRF | Long-Short-Term-Memory-Conditional-Random-Fields Classifiers |
| NLP | Natural Language Processing |
| NLTK | Natural Language Tool Kit |
| ML | Machine Learning |
| MR | Meta Requirements |
| OECD | Organisation for Economic Co-operation and Development |
| PASL | Perceived Argumentation Skill Learning |
| PFA | Perceived Feedback Accuracy |
| PU | Perceived Usefulness |
| POS | Part-of-Speech |
| RQ | Research Question |
| SVM | Support Vector Machine |
| SD | Standard Derivation |
| TG | Treatment Group |
| TML | Technology-Mediated Learning |
| US | User Stories |

# Abstract

Argumentation is an omnipresent rudiment of daily communication and thinking. The ability to form convincing arguments is not only fundamental to persuading an audience of novel ideas but also plays a major role in strategic decision-making, negotiation, and constructive civil discourse. However, humans often struggle to develop argumentation skills owing to a lack of individual and instant feedback in their learning process. Providing feedback on the individual argumentation skills of learners is time-consuming and not scalable if conducted manually by educators. With the help of new adaptive argumentation learning systems, students would be able to learn autonomously and independently of the instructor, time, and place. Although the use of Artificial Intelligence (AI) seems to be a promising approach, current literature lacks (1) insights on the theory-based and learner-centered requirements for adaptive argumentation skill learning, (2) knowledge in how to design and embed AI-based adaptive argumentation learning systems to improve students' learning outcomes and user-experience, and (3) insights on the influence of adaptive argumentation learning systems on students argumentation skills.

This dissertation tackles these literature gaps to explore the potential of adaptive argumentation skill learning based on AI by designing, implementing, and evaluating new technology-enhanced pedagogical concepts to actively support students in developing the ability to argue in a structured, logical, and reflective way.

Using a design science research approach, I develop new student-centered pedagogical scenarios with empirically evaluated design principles, linguistic corpora, ML algorithms, and innovative learning tools. I introduce and contribute to two novel classes of technology-mediated argumentation learning tools: (1) AI-based argumentation writing support system and (2) conversational argumentation learning systems. My results indicate that the two novel classes of adaptive learning tools based on ML algorithms and user-centered design patterns help students to develop better argumentation writing skills in different domains. Thereby, I firstly bridge the boundaries of argumentation learning and argumentation mining by examining new pedagogical scenarios for adaptive argumentation learning from a student-centered perspective. Therefore, I contribute not only with new rich argumentation annotation schemes, argumentation corpora, and novel ML models in German language, but most of all insights into the overall socio-technical embedding, design, and their effects of AI-based argumentation writing support systems and conversational argumentation learning systems to teach students how to argue.

Keywords: Adaptive Argumentation Learning; Pedagogical Conversational Agents; Argumentation Mining; Technology-mediated Learning; Design Science Research.

# Zusammenfassung

Die Fähigkeit, überzeugende Argumente zu formulieren, ist nicht nur von grundlegender Bedeutung, um ein Publikum von neuen Ideen zu überzeugen, sondern spielt auch eine wichtige Rolle bei strategischen Entscheidungsfindungen, bei Verhandlungen und im allgemeinen politischen Diskurs. Menschen haben jedoch oft Schwierigkeiten, Argumentationsfähigkeiten zu entwickeln. Ein Grund dafür ist der Mangel an individuellem und formativem Feedback im Lernprozess von Studierenden oder in der Aus-und Weiterbildung. Eine Rückmeldung über den Stand der individuellen Argumentationsfähigkeiten von Lernenden ist zeitaufwändig und für Lehrkräfte nicht skalierbar. Neuartige adaptive Argumentationslernsysteme haben das Potential, Studierende zu unterstützen selbstständig und unabhängig von Lehrkräften, Zeit und Ort zu lernen. Obwohl der Einsatz von Künstlicher Intelligenz (KI) ein vielversprechender Ansatz zu sein scheint, fehlt es in der aktuellen Literatur (1) an Erkenntnissen über die theoriebasierten und lernerzentrierten Anforderungen an Lernen von Argumentationsfähigkeiten mit adaptiven Lerntools, (2) an Wissen darüber wie KI-basierte adaptive Argumentationslernsysteme konzipiert und gebaut werden können, um die individuellen Lernpfade und die Nutzung der Studierenden zu verbessern und (3) an Erkenntnissen über den Einfluss adaptiver Argumentationslernsysteme auf die eigentlichen Argumentationsfähigkeiten der Studierenden.

Diese Dissertation greift diese drei Forschungslücken auf und untersucht das Potential von adaptivem Lernen von Argumentationsfähigkeiten mit Hilfe von KI. Dazu werden neue technologiegestützte pädagogische Konzepte entworfen, implementiert und evaluiert, die Studierende aktiv unterstützen, strukturiert, logisch und reflektiert zu argumentieren. Auf Basis eines designwissenschaftlichen Forschungsansatzes entwickle ich neue studierendenzentrierte pädagogische Szenarien mit empirisch evaluierten Designprinzipien, linguistischen Korpora, ML-Algorithmen und innovativen Lernwerkzeugen. Ich stelle dazu zwei neuartige Klassen von IT-basierten Lernwerkzeugen für das Argumentieren vor: (1) KI-basierte Systeme zur Unterstützung des Schreibens von Argumenten und (2) dialogbasierte Argumentationslernsysteme. Meine Ergebnisse zeigen, dass die beiden neuen Systemklassen den Studierenden helfen bessere Argumentationsfähigkeiten in verschiedenen pädagogischen Bereichen zu entwickeln. Dabei überbrücke ich erstmals die Grenzen von Argumentationslernen und Argumentation Mining, indem ich neue pädagogische Szenarien für adaptives Argumentationslernen aus einer studierendenzentrierten Perspektive untersuche. Daher trägt diese Arbeit nicht nur mit neuen, reichhaltigen Argumentationsannotationsschemata, Argumentationskorpora und neuartigen ML-Modellen in deutscher Sprache bei, sondern vor allem mit Einblicken in die allgemeine sozio-technische Einbettung, das Design und die Auswirkungen von KI-basierten Argumentationslernsysteme, um Lernenden zu helfen besser zu argumentieren.

# 1 Introduction

> *"We need technology in every classroom and in every student and teacher's hand, because it is the pen and paper of our time, and it is the lens through which we experience much of our world."* – David Warlick

## 1.1 Problem Statement

In recent years, the educational landscape has changed dramatically. Multiple learning formats have transformed from traditional offline learning scenarios to large-scale online formats such as massive open online courses (Noetel et al., 2021; Xu et al., 2021). Individual support and personal recommendations for students in their learning process are still pendent challenges and have not yet been solved in most of these pedagogical scenarios (e.g., Kulik and Fletcher (2016); Vygotsky (1980); Noetel et al. (2021)). Due to financial and organizational constraints, educational institutions such as high schools and universities struggle to provide individual help. Large classroom sizes in public schools, mass lectures with more than 100 students per professor at universities, and the current rise in massive open online courses with more than 1000 participants all obstruct individual interaction between students and instructors (Seaman et al., 2018). This effect has been intensified by the COVID-19 pandemic in 2020/21 and the accompanying governmental lockdowns. Consequently, technology-mediated distance-learning scenarios have become a reality for many students and educators. In 2017, 33.1% of learners around the world took at least one course online, compared to 24.8% in 2012 (Lederman, 2018). With the leading MOOC provider Coursera's launch of a successful initial public offering in March 2021, further expansion of distance learning settings is predicted (Tse and Roof, 2021).

Not only the way we learn is changing rapidly, but also what learning content is taught is subject to an ongoing transformation due to significant technological advances in our societies (i.e., Noetel et al. (2021); Xu et al. (2021)). For example, compared to a few decades ago, information is readily available in today's world, and the ability to memorize existing information is no longer prioritized (Oecd, 2018, 2019; WEF, 2018). On the contrary, the skills and abilities required to structure and process knowledge are gaining importance (WEF, 2018). Job profiles have thus shifted toward interdisciplinary, ambiguous, and creative tasks (vom Brocke et al., 2018). Educational institutions must not only adapt to the above-mentioned transformation towards online learning but also evolve their curricula when it comes to the composition of skills and knowledge conveyed (Topi, 2018). Critical thinking, teamwork, and problem-solving are examples of metacognition skills that have become increasingly crucial (Fadel et al., 2015). International organiza-

tions such as the Organization for Economic Cooperation and Development (OECD) and the World Economic Forum (WEF) have urged for a shift in student education to prepare future workforces with the necessary skills (WEF, 2018). In fact, the OECD included metacognition skills as a major element in their learning framework 2030 (Oecd, 2018). One example of these competencies is the ability to advance structured, reflective, and well-formed arguments (Toulmin, 2003). Argumentation is not only essential to daily communication and thinking but also contributes significantly to multiple metacognition skills, such as communication, collaboration, and problem-solving (Kuhn, 1992). In studies dating back to Aristotle, the ability to form convincing arguments has been recognized as essential to persuading audiences of novel ideas' merits and as playing major roles in analyzing different standpoints and in constructive democratic civil discourse (e.g., for citizens to assess whether certain news items are fake, Diana (2018)).

However, teaching metacognition skills such as argumentation in professional organizations or educational institutions always has been and still is hindered by a lack of individual and personal learning experiences (i.e., Hattie and Timperley (2007); Vygotsky (1980)). Learners are frequently confronted with unfavorable educator-student ratios, not only in the distance-learning scenarios of massive open online courses (Seaman et al., 2018) but also in traditional large-scale lectures on university campuses. Hence, information systems (IS) have been leveraged as technology-mediated learning tools over several decades to meet scalable and individual learning challenges (Alavi and Leidner, 2001; Gupta and Bostrom, 2009; Leidner and Jarvenpaa, 1995). IS have been designed and developed to support individuals in learning how to argue in online debates (Wang et al., 2020) or collaborative learning settings (Dillenbourg et al., 2009) by receiving tailored argumentation feedback irrespective of instructor, time, and place (Afrin et al., 2021; Kulik and Fletcher, 2016; Wang et al., 2020). In fact, for at least 35 years (i.e., Smolensky et al. (1988)), researchers have developed technology-enhanced argumentation support systems in various domains, including law (Pinkwart et al., 2009), science (Osborne et al., 2016; Suthers and Hundhausen, 2001), and conversational argumentation (De Groot et al., 2007). Nevertheless, the design and adoption of current argumentation learning systems lags behind recent technological developments in IS, human—computer interaction, natural language processing (NLP) research and machine learning (ML). To unleash the full potential of NLP and ML for adaptive argumentation skill learning, it is necessary to better understand how the technology can support students' individual learning processes in order to address the problem of a lack of individual support in large-scale and distance learning scenarios (Noetel et al., 2021; Xu et al., 2021). Hence, with this thesis, I identified three overarching research challenges that need to be addressed. In the following, I will shortly introduce the three research challenges for AI-based argumentation learning systems. Afterwards, I will propose research questions (RQ) that aim to address the

distinct challenges.

First of all, past research lacks interdisciplinary literature and design knowledge to build accessible, intelligent, and learner-centered argumentation tools that enable self-regulated learning (Huang et al., 2021; Noetel et al., 2021; Xu et al., 2021). Therefore, it is not only learners in IS education who still struggle to develop argumentation skills owing to a lack of adaptive and instant monitoring and evaluation in their individual learning process, since current approaches fail to provide scalable and formative feedback on learners' argumentation skills (Lawrence and Reed, 2019; Scheuer, 2015; Pinkwart et al., 2009). In fact, a socio-technical IS perspective to support the formation of a holistic learner-centered argumentation tutoring tool that provides individuals with the opportunity to monitor and evaluate themselves based on advances in NLP and ML is missing (Huang et al., 2016; Lawrence and Reed, 2019; Scheuer, 2015; Scheuer et al., 2010).

**Research Challenge 1:** *There is limited knowledge about the technological embedding of novel advances from NLP and ML for adaptive argumentation learning systems as well as about the use and the corresponding requirements for the design of novel learning tools for argumentation skill learning.*

Despite the existence of literature on argumentation systems, insights into overall system design and development—particularly for new AI-based argumentation learning systems—are rare. Furthermore, a general conceptual approach to technology-based argumentation skill learning is lacking. Research that considers not only the educational task (writing coherent arguments) but also the design of holistic argumentation learning systems embedded in pedagogical scenarios (structure) and focuses on the learner through a user-centered and theory-motivated design is scant (e.g., Bostrom and Heinen (1977); Gupta and Bostrom (2009)), which leads to research challenge 2.

**Research Challenge 2**: *There is limited knowledge on how to design and embed adaptive argumentation learning systems to improve students' learning outcomes and learner experience.*

Although existing studies have extended the understanding of technology-mediated argumentation learning, they usually sought to address specific research questions rather than gather information for widespread application (e.g., Afrin et al. (2021); Chernodub et al. (2019); Lippi and Torroni (2016a); Wang et al. (2020); Huang et al. (2021); Noetel et al. (2021); Xu et al. (2021)). While earlier studies made contributions towards exploring the development and use of argumentation learning systems, they often neglected presenting

the general design principles and implementations—particularly for the novel class of AI-based argumentation learning systems—for further developments in the field. To the best of my knowledge, by the time of starting this thesis, few research has been published on technology-mediated argumentation learning from IS perspectives, and no studies have synthesized and conceptualized design principles for the novel class of AI-based argumentation skill learning. Specifically, empirical evaluations about the embedding of a learner-centered design for argumentation learning systems are rare. In the literature, there exists no approach with principles, design knowledge, and evaluations on how to design and embed an adaptive argumentation learning system in a pedagogical scenario to help students learn how to argue.

As Scheuer (2015), p. 126 identifies, "rigorous empirical research with respect to adaptation strategies is almost absent; a broad and solid theoretical underpinning or theory of adaptation for collaborative and argumentative learning is still lacking". Besides, novel interaction paradigms such as pedagogical conversational agents promise to provide new forms of individual guidance and feedback to students, for instance, when completing a task to write persuasive texts through a natural dialogue-based interface. However, past research does not investigate these novel interaction modalities for the design of AI-based argumentation learning systems. Hence, the following third research challenges is derived:

**Research Challenge 3**: *There are limited insights and empirical investigations on the influence of adaptive argumentation learning systems on students' argumentation skills.*

## 1.2 Solution Statement and Research Questions

The main objective of the proposed dissertation project is to contribute to the three outlined research challenges in the fields of NLP, human—computer interaction, and IS research. To fulfill this overall objective, my thesis concerns the investigation of two overarching research areas, which relate to (1) *how to design novel argumentation learning systems based on recent advances in AI to support argumentation skill learning independent of an instructor, time, and location* and (2) *how these design findings influence students' argumentation skills and learner experience.*

I organize my thesis and the corresponding research questions according to the conceptual model of technology-mediated learning by Gupta and Bostrom (2009). The framework guides technology-mediated learning research based on three different dimensions: (1) *learning context*, (2) *learning method structures*, and (3) *learning process*. Figure 1 illustrates the technology-mediated learning model in combination with the identified chal-

lenges. With my dissertation, I focus on a constructivist learning perspective (Vygotsky, 1980) to enhance metacognition skill learning (the learning goal) in distance learning environments (learning context). The learning method structures are built upon recent advances in AI to better guide and support argumentation skill learning processes with argumentation learning systems. Finally, the goal is to investigate the impact of argumentation learning systems based on AI on the learning process and resulting learning outcomes of students.

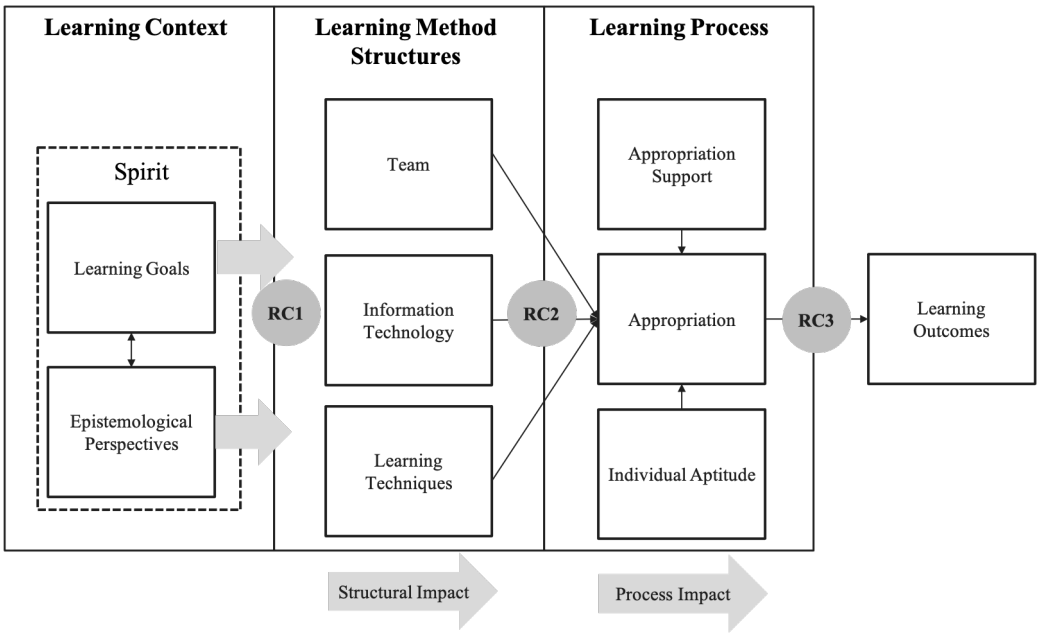



Figure 1: Focus of the dissertation based on the conceptual model of technology-mediated learning by Gupta and Bostrom (2009).
Source: Gupta and Bostrom (2009)

The first research question focuses on the technical foundations of NLP and ML for adaptive learning tools and the requirements for argumentation learning systems derived from the literature and users to tackle research challenge 1. The second research question focuses on how to design argumentation learning systems to improve students' argumentation skill learning independent of an instructor, time, and location (research challenge 2). The third research question investigates to what extent an AI-based adaptive learning tool influences the argumentation skills of students (research challenge 3).

**RQ1:** *What are the foundations and requirements for adaptive learning tools to teach students how to argue?*

The dissertation starts with the foundations and an understanding of past research in NLP and ML on creating adaptive user-interaction tools. This is necessary since the design of

an AI-based argumentation learning system is a complicated and complex endeavor that builds upon recent findings in different disciplines such as NLP, IS, and human—computer interaction. To better structure RQ1, I divide the research question into two sub-research questions:

**RQ1a:** *What are the theoretically grounded and empirically validated dimensions and characteristics of NLP features for ML-based IS?*

**Method:** *Systematic literature review, taxonomy development, and technical experiments.*

**Results:** *A taxonomy of NLP features for ML-based IS and a transfer learning approach for argumentation identification across text domains.*

Accordingly, I start by conducting a systematic literature review to structure NLP features for ML-based IS in the form of a taxonomy based on the process of Nickerson et al. (2013), and I investigate the potential of novel transformer-based transfer learning approaches for argumentation mining. Both the corpus-diagnostic algorithm and the NLP feature taxonomy assist researchers and designers to build artifacts based on argumentation mining, e.g., adaptive argumentation feedback tools. The NLP feature taxonomy and the algorithm provide a solid theoretical and practical foundation to further impact the design of adaptive learning tools for argumentation skills.

**RQ1b:** *What are user and literature requirements for argumentation learning system to support students in their argumentation learning process?*

**Method:** *Systematic literature review from an interdisciplinary perspective and user interviews.*

**Results:** *Literature issues and user requirements for the design of an argumentation learning system.*

In RQ1b, my aim is to develop a holistic understanding of the problem covered by my dissertation. I do this by investigating the state of the art in practice and literature on adaptive argumentation learning systems and by formulating design requirements for a possible solution. To investigate scientific literature about argumentation learning systems, I conduct a systematic literature review following the methodological approaches of Cooper (1988); Webster and Watson (2002); vom Brocke et al. (2015). To receive further insights from learners, I conduct 30 semi-structured interviews with potential users of an

adaptive argumentation learning tool to include potential requirements from an end-user perspective. I interview students following the expert interview method by Glaser and Laudel (2010) to get a solid understanding of their needs for adaptive skill and argumentation learning.

**RQ2:** *How can an adaptive learning tool be designed based on AI in a large-scale pedagogical scenario to support the development of students' argumentation skills in their learning process?*

Next, I investigate how to design different pedagogical scenarios in which an adaptive learning tool can help students to learn how to argue through adaptive argumentation tutoring and support in their writing process. To do so, I use methods from human-centered design research, and conduct annotation studies and technical experiments to model argumentation discourse structure in German student-written texts. I rely on my findings from RQ1 to formulate user stories (based on the semi-structured interviews) and meta-requirements (based on the identified scientific literature). The user stories and the meta-requirements help me to formulate initial design principles (based on Gregor et al. (2020)). The design principles are intended to guide and assist designers in developing new argumentation learning tools in different pedagogical scenarios to impact the development of students' argumentation skills. Based on the initial design principles I derive certain design features to develop first mock-up-based scenarios. I test these low-fidelity design artifacts with end users to further refine my design principles and finally formulate a concise but comprehensive set of preliminary design principles. Moreover, I investigate the technical foundations for the design of an adaptive argumentation learning tool based on artificial intelligence (AI). To do so, the availability of labeled corpora is crucial for designing, training, and evaluating argumentation mining algorithms. However, at the start of my thesis, no corpora were available for argumentation mining based on German student-written texts. Therefore, I developed two novel annotation schemes to capture argumentation relations and argumentation discourse of student-written texts in German to create two corpora to design argumentation mining models for adaptive argumentation learning.

**RQ3:** *To what extent does an AI-based adaptive learning tool influence the argumentation skills of students in different pedagogical scenarios?*

For RQ3, I validate the hypothesis that individual tutoring and individual feedback from an adaptive learning tool based on AI on a students' argumentation skills lead to better

argumentation skills of the students. I investigate to what extent adaptive argumentation learning systems impact the user perception and the argumentation skills of a student as well as what characteristics are important to do so in a series of laboratory and field experiments. Hence, I instantiate the design principles from RQ2 as design features in different argumentation learning artifacts. I build the adaptive logic of my artifact on the foundations of my developed algorithm, the NLP feature taxonomy, the annotation scheme, and the argumentation-annotated corpus from RQ2. In this vein, I introduce and contribute to two novel classes of argumentation learning systems: AI-based argumentation writing support systems and conversational argumentation learning systems. RQ3 is divided into three research sub-questions (RQ3a, RQ3b, RQ3C), each representing a different pedagogical scenario with an instantiation and evaluation of one of the two classes (argumentation writing support systems and conversational argumentation learning systems):

**RQ3a:** *How does an adaptive argumentation writing support system need to be designed to influence students' argumentation skills in peer review scenarios?*

**Method:** *Design science research including corpus development, argumentation modeling in student-written peer reviews, and mixed-method research with a series of experiments.*

**Results:** *Empirical evidence for the influence of an adaptive argumentation writing support system based on certain design principles on students' formal metacognitive argumentation skills and their user experience.*

The first scenario consists of a constructive teaching scenario. Here, I provide students with adaptive feedback based on their written text by highlighting potential pitfalls in their argumentation based on a static interaction with adaptive and individual argumentation feedback. To test if individual feedback on students' argumentation helps them to write more convincing texts and thus increase their short-term argumentation skills, I compare the pedagogical scenario with an instantiated argumentation writing support system to a proven argumentation support tool based on an input mask (Fischer et al., 2013). Second, I investigate if a constructive pedagogical scenario with adaptive feedback increases the long-term argumentation skills of students. Hence, I instantiate an argumentation writing support system that provides students with adaptive argumentative feedback on written peer feedback about business models. This artifact is tested in a large-scale lecture as a field experiment. After a period of two months and nine treatments, I then evaluate if the argumentation skills of the treatment group are higher in a second argumentation domain (societal debates based on Flender et al. (1999)) than of the control group, which just received general argumentation feedback on their written

peer feedback.

**RQ3b:** *What requirements should be considered when designing an adaptive conversational argumentation learning system that aims to support students in writing argumentative texts?*

**Method:** *Design science research including a series of experiments.*

**Results:** *Empirical evidence of the impact of conversational argumentation writing support system based on certain design principles on students' argumentation skills and their learning experience.*

In RQ3b I investigate if an interactive pedagogical scenario (Chi and Wylie, 2014) based on a dialogue system (e.g., a conversational agent) helps students learn how to argue through individual tutoring, tasks scaffolding, and theory input. My goal is to provide students with assistance comparable to a discussion with a human tutor. Therefore, I instantiate a conversational argumentation learning systems with which I validate the hypothesis that an interactive dialogue-based system with adaptive argumentation feedback and tutoring helps student to learn how to argue (Chi and Wylie, 2014). I validate the artifacts and their influence on students' abilities to argue in two different experiments. After that, I describe my findings in a nascent design theory with reference to Gregor's anatomy of a design theory (Jones and Gregor, 2007).

**RQ3c:** *How can an argumentation writing support system be designed based on nudging and self-evaluation to influence students' argumentation writing skills?*

**Method:** *Design science research including corpus development, argumentation modeling in student-written business pitches, and mixed-method research with a series of experiments.*

**Results:** *Further empirical evidence for the influence of argumentation writing support systems based on certain design principles on students' argumentation skills and their learning experience.*

In order to consider learner centricity, I draw on the concept of nudging (RQ3c), which is originally a concept from behavioral economics (Thaler and Sunstein, 2009) to nudge users to utilize computational methods in their learning process. Embedding nudges in traditional learning environments, but also digital nudges in learning tools, has shown

positives effects on learning outcomes, especially for self-regulated learning such as self-monitoring and self-evaluation (Caraban et al., 2019; Hartwig and Reuter, 2021; Weijers et al., 2020). To investigate how individual argumentation self-evaluation will help students write more convincing texts, I designed an argumentation writing support system called ArgumentFeedback based on nudging theory and evaluated it in a series of three studies with a total of 108 students. I found that students who receive a self-evaluation nudge wrote more convincing texts with a better quality of argumentation compared to two alternative groups. The measured self-efficacy and the technology acceptance provide promising results for embedding adaptive argumentation writing support tools in combination with digital nudging in traditional learning settings to foster self-regulated learning.

By formulating the three overarching research questions and answering them in the course of my dissertation, I provide several contributions to adaptive argumentation learning research, particularly in the domains of human—computer interaction, NLP, IS, and education research.

The dissertation's overall purpose is to address the research challenges presented in subsection 1.1, and achieving this is the study's major contribution to research in this area. To summarize, the dissertation offers three major theoretical contributions to the fields of IS and education. First, by leveraging recent advances in argumentation mining, I contribute to human—computer interaction and IS research by designing, building, and evaluating new forms of adaptive argumentation learning tools that provide students with feedback on the argumentation structure of a text independent of an instructor, time, or location. To establish a clear set of design principles for adaptive argumentation learning tools, I rigorously derived requirements from scientific articles and 30 semi-structured user interviews. I refined and evaluated those design principles resulting in instantiated two novel system classes (AI-based argumentation writing support systems and conversational argumentation learning systems) in different experiments. I tested the impact of the design principles for adaptive argumentation learning tools through these artifacts in three pedagogical scenarios. I found that students that used an AI-based argumentation writing support system (e.g., my systems ArgueLearn or ArgumentFeedback) or a conversational argumentation learning system (e.g., my system ArgueTutor) improved their argumentation skills significantly compared to students who only received general or no writing support.

Second, aside from the software artifact as a real-world application of an adaptive argumentation learning tool, I will contribute a nascent design theory for the design of AI-based argumentation writing support systems and a theory for the design of conversational argumentation learning systems to the scientific knowledge base. Due to the

systematic procedure of my thesis, I generate a satisfying design contribution (Jones and Gregor, 2007).The generated design knowledge is not just applicable to a single use case, but it may also be applied to other cases in argumentation learning systems. For instance, it is easily possible to apply the concept of my software artifacts (e.g., ArgueLearn or ArgueTutor) in courses that deal with other content than business models or use other languages. Only the back-end algorithm needs to be modified to a different environment for this purpose. In the literature, there are a variety of corpora and argumentation mining models that can be simply included into ArgueLearn, such as for English student writings (Stab and Gurevych, 2017) or English law cases (Mochales and Moens, 2008). These use cases do not necessitate any changes to the form and function design principles or the overall system design. Furthermore, the design expertise can be applied to instructional situations that focus on the development of different metacognition skills or language domains. A comparable adaptive writing support system can be utilized, for example, if students' general feedback abilities or empathy skills are to be trained. However, in this scenario, the system architecture may need to be partially updated, such as the graph visualization or argumentation dashboard.

Third, I also contribute to NLP research. In the recent decade, research in argumentation learning and mining has made significant progress. We understood a lot about the mechanisms, potential benefits, and flaws in both study streams. However, both study domains had not yet crossed their borders to establish a full and multidisciplinary relationship between argumentation learning for students and argumentation mining technology when I began this dissertation. I firstly bridged these boundaries by examining a pedagogical scenario for adaptive argumentation learning from different perspectives, i.e., human—computer interaction, IS, and NLP. As a result, I offer not only new rich argumentation annotation schemes, argumentation corpora, and novel machine learning models, but also insights into the overall embedding and design of AI-based argumentation writing support systems, as well as the potential of dialogue-based argumentation learning, such as conversational argumentation learning systems. These findings offer a new perspective at AI-based adaptive argumentation and more general metacognition skill learning that has been overlooked in previous studies.

In addition to contributing to the scientific knowledge base, the findings are useful for the practical use of AI techniques such as NLP and ML-based feedback systems for education. My findings show that current state-of-the-art NLP and ML techniques are well-suited to designing complex adaptive skill learning systems that can assist individuals in learning new abilities — even in existing pedagogical contexts like peer evaluations. As a result, I've demonstrated that, in contrast to present e-learning methodologies, advanced assistance capabilities in skill education may be provided utilizing NLP and ML techniques.

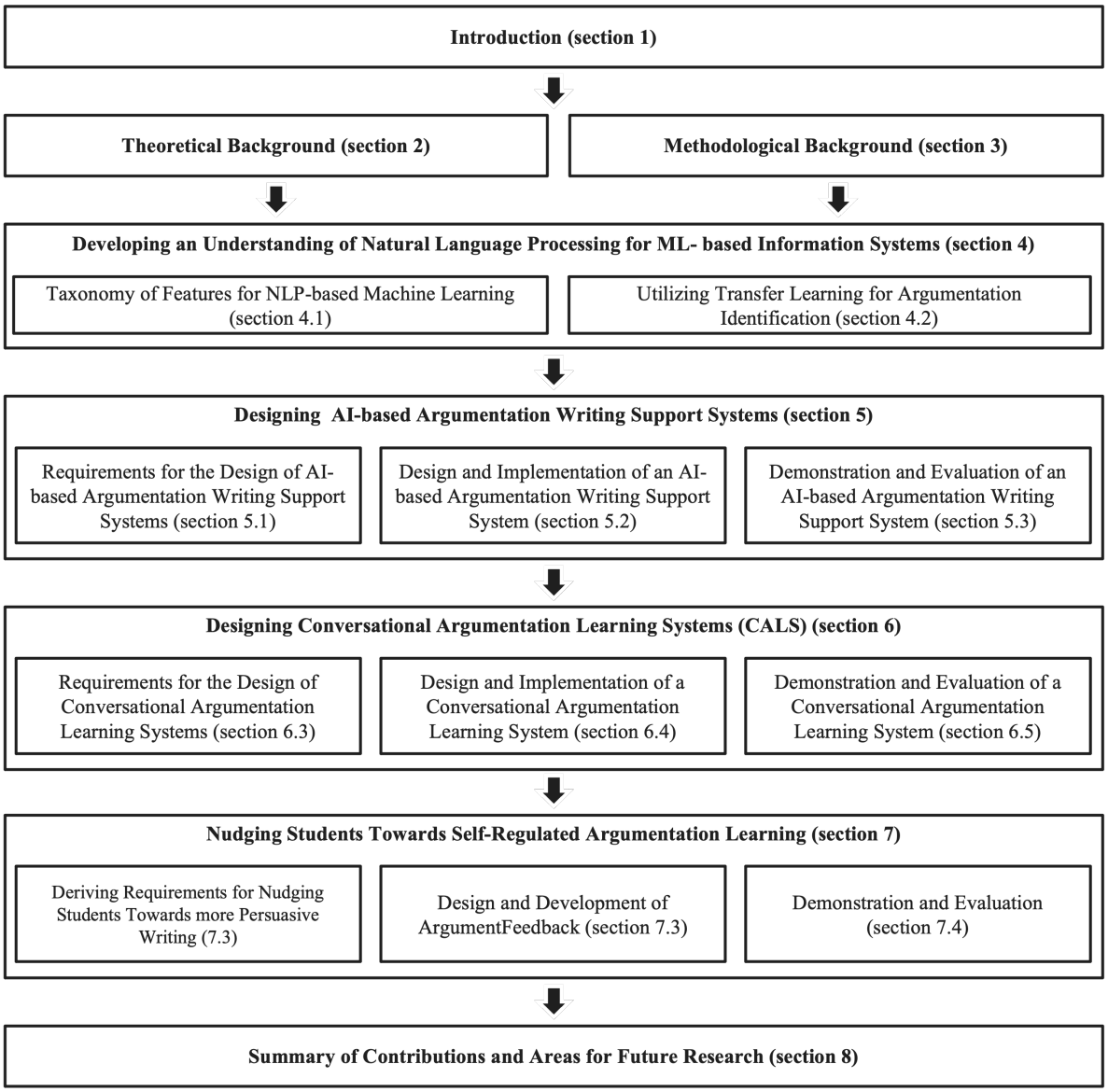



Figure 2: Overview of the dissertation structure in eight consecutive sections. Source: Own illustration

## 1.3 Structure of the Dissertation

To achieve the outlined objective of this dissertation project and to answer the stated research questions presented in subsection 1.2, I structure my dissertation in the subsequent chapters as outlined in Figure 2.

I will start by providing the reader with the necessary theoretical and conceptual background concerning technology-mediated learning, argumentation skill learning, past argumentation learning systems, and the background on the technical foundations (NLP, ML, argumentation mining) of my work. Next, I briefly sketch an overview of the methodological background and the applied qualitative and quantitative methods in section 3. In

section 4, I focus on proving a structured overview of NLP-based features for ML-based IS. Using a systematic literature review and the taxonomy development process of Nickerson et al. (2013), I structure NLP features for the efficient design of intelligent IS. Moreover, I also outline a technical experiment to investigate the potential of transformer-based transfer learning for argumentation identification as a basis for the technical implementation of my thesis. Equipped with the technical basics, section 5 investigates the major challenges of adaptive argumentation skill learning research that are described in subsection 1.1 to guide the further course of the dissertation. Moreover, I derive requirements from potential end users and the literature. Based on the requirements, I formulate design principles for AI-based argumentation writing support systems and evaluate them in an instantiation for the pedagogical scenario of peer reviews in two studies. In order to do so, I derive a new annotation scheme for argumentative student peer reviews and trained a model to detect the argumentation structures in these texts. I instantiate a novel AI-based argumentation writing support system called ArgueLearn and evaluate it in two empirical studies.

In section 6, I follow a similar design science approach and use the corpus and the model from section 5 to design and evaluate a conversational argumentation writing tutor for the domain of peer reviews. After the evaluation and the better understanding of adaptive argumentation learning based on AI, I connect and evaluate my findings to nudge students towards self-regulated argumentation learning in the pedagogical scenario of persuasive elevator pitches in section 7. To do so, I transfer my annotation scheme from section 5 to the domain of student business pitches and annotate and evaluate a new argumentative corpus for student-written business pitches.

Section 8 summarizes my design finding. This is followed by a summary of the dissertations' contributions from a theoretical and practical perspective along with recommendations for future research avenues for skill learning based on NLP and ML.

## 1.4 Overview of Publications in Relation to the Research Questions

My thesis is based on and related to 13 scientific publications that contributed to answering the stated research questions. Table 1 illustrates an overview of past publications that addressed the proposed research questions of this dissertation. Additionally, I provide information regarding outlet metrics (if available) such as the Google Scholar h5-index (h5)[1] and the ranking of VHB JOURQUAL 3 (JQ3)[2]. Furthermore, Table 1 sketches an

[1] https:/scholar.google.com/citations?view$_op = top_venueshl = de$

[2] https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3/gesamtliste

overview of the predominant methodological focus as well as the research context of the publications. At the beginning of each section of my dissertation, I also highlight how these publications have contributed to the content.

| **No.** | **Publication** | **Outlet / Metrics** | **Method and Context** | **RQ** |
|---|---|---|---|---|
| 1 | Hansjörg Fromm, Thiemo Wambsganss, and Matthias Söllner. 2019. Towards a Taxonomy of Text Mining Features. In *In Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm, Sweden | H5: 34 VHB JQ3: B | Literature Review; Taxonomy Development | RQ1a |
| 2 | Thiemo Wambsganss, Christian Engel, and Hansjörg Fromm. 2021a. Improving Explainability and Accuracy through Feature Engineering : A Taxonomy of Features in NLP-based Machine Learning. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas | H5: 34 VHB JQ3: A | Literature Review; Taxonomy Development | RQ1a |
| 3 | Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020a. Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach. In *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany | H5: 22 VHB JQ3: C | Literature Review; Technical Experiment | RQ1a |
| 4 | Thiemo Wambsganss and Roman Rietsche. 2019. Towards designing an adaptive argumentation learning tool. In *40th International Conference on Information Systems, ICIS 2019*, Munich, Germany | H5: 34 VHB JQ3: A | Design science; Deductive; Literature Review; User Interviews | RQ1b |

| No. | Publication | Outlet / Metrics | Method and Context | RQ |
|---|---|---|---|---|
| 5 | Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020b. AL : An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA. Association for Computing Machinery | H5: 101 VHB JQ3: / | Design Science; Deductive; Higher education; Distance learning environment | RQ2/ RQ3a |
| 6 | Thiemo Wambsganss, Tobias Kueng, Matthias Söllner, and Jan Marco Leimeister. 2021c. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA | H5: 101 VHB JQ3: / | Design Science; Deductive; Higher education; Distance learning environment | RQ2/ RQ3b |
| 7 | Thiemo Wambsganss. 2021. Designing Adaptive Argumentation Learning Systems Based on Artificial Intelligence. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* | H5: 101 VHB JQ3: / | Design Science; Deductive; Higher education | RQ2/ RQ3a |
| 8 | Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020c. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*, Barcelona, Spain | H5: 64 VHB JQ3: / | Corpus development; Inductive; Technical experiments | RQ3a |
| 9 | Thiemo Wambsganss, Matthias Söllner, Andreas Janson, Ken Koedinger, and Jan Marco Leimeister. 2022b. Improving Argumentation Based on a Dynamic Feedback System Created with Machine Learning. *Under review at Information Systems Research*, pages 1–38 | H5: 50 VHB JQ3: A+ | Inductive and Deductive; Quantitative experiments | RQ3a |

| No. | Publication | Outlet / Metrics | Method and Context | RQ |
|---|---|---|---|---|
| 10 | Thiemo Wambsganss, Matthias Söllner, and Jan Marco Leimeister. 2020d. Design and Evaluation of an Adaptive Dialog-Based Tutoring System for Argumentation Skills. In *International Conference on Information Systems (ICIS)*, Hyderabad, India | H5: 34<br>VHB JQ3: A | Design Science; Deductive and Inductive; Higher education | RQ3b |
| 11 | Thiemo Wambsganss, Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. 2021b. AI-based Argumentation Tutoring – A Novel System Class to Improve Learners' Argumentation Skills. In *Annual Meeting of the Academy of Management (AOM). - Virtual Conference, 2021*, volume 2021, page 10233 | H5: 4<br>VHB JQ3: / | Design Science; Deductive and Inductive; Higher education; Distance learning environment | RQ2/ RQ3a |
| 12 | Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760, Dublin, Ireland. Association for Computational Linguistics | H5: 157<br>VHB JQ3: / | Corpus development; Inductive; Technical experiments | RQ3c |
| 13 | Thiemo Wambsganss. 2022. Nudging Towards Self-Regulated Argumentation Learning: Designing Intelligent Argumentation Writing Support Systems. *Institute of Information Management at the University of St.Gallen Working Paper, St. Gallen, Switzerland*, pages 1–30 | H5: /<br>VHB JQ3: / | Design Science; Deductive; Higher education; Distance learning environment | RQ3c |

Table 1: Overview of publications related to the dissertation according to the research questions.
Source: Own illustration

# 2 Theoretical Background

> *"It is in argumentation that we are likely to find the most significant way in which higher order thinking and reasoning figure in the lives of most people. Thinking as argument is implicated in all of the beliefs people hold, the judgments they make, and the conclusions they come to; it arises every time a significant decision must be made. Hence, argumentative thinking lies at the heart of what we should be concerned about in examining how, and how well, people think"*
>
> - Kuhn (1992), pp. 156–157.

As stated by Kuhn (1992) argumentation plays a pivotal role in our communication, in our thinking, and therefore in our daily lives. To better understand how argumentation skill learning can be supported by IT, I aim to provide the reader a theoretical background on technology-mediated argumentation learning. Based on the paradigm of Gupta and Bostrom (2009), I will briefly introduce the technology-mediated learning framework as well as the three dimensions of learning context, learning method structures, and learning process in subsection 2.1. Afterwards, I dive into each of these dimensions and outline how the concepts are defined with regards to argumentation learning. Therefore, in subsection 2.2 I will outline the learning context of argumentation skill learning and the underlying constructivist learning paradigm I aim to follow in this dissertation. Next, in subsection 2.3 I will provide a deep dive into the different learning method structures and the different system classes that have been developed in the last decades for technology-mediated argumentation learning. Here, I will also introduce and define two new classes, namely AI-based argumentation writing support systems and conversational argumentation learning systems. Third, I will briefly state the difference between "arguing to learn" and "learning to argue" for the learning process in argumentation learning in subsection 2.4. Last but not least, since my dissertation follows an interdisciplinary approach based on methods from human—computer interaction, NLP, and IS, subsection 2.5 provides a short overview of NLP, ML, and argumentation mining to ensure the reader is equipped with the necessary background knowledge of the fundamental technologies applied in this dissertation.

## 2.1 Technology-Mediated Learning

The objective of this dissertation is to investigate (1) how to design novel argumentation learning systems based on recent advantages in AI to support argumentation skill learning

independent of an instructor, time, and location, and (2) to answer how these design findings influence students' argumentation skills and learning experience. Hence, the focus of this work is mainly on the domain of digital learning environments, since they are typically embedded around large-scale and distance learning scenarios.

In IS research, the learning embedded around technology-mediated learning forms a prominent literature stream (Gupta and Bostrom, 2009; Alavi and Leidner, 2001). Alavi and Leidner (2001) defined technology-mediated learning as *"an environment in which the learner's interactions with learning materials, peers, and/or instructors are mediated through advanced information technology"* (Alavi and Leidner (2001) p.2). As Gupta and Bostrom (2009) noted, technology-mediated learning includes "all the elements of a social-technical system: technology and learning techniques, process, actors, actions, and outcomes" (Gupta and Bostrom (2009), p.3). Scholars in different disciplines have used the terms e-learning, online learning, distance learning, technology-enhanced learning, and IT-supported learning synonymously (Gupta and Bostrom, 2009), so I will do the same in this dissertation.

To better embed my research in related literature, I used the theoretical framework of Gupta and Bostrom (2009) as a guide for the structure of my work. The framework defines both the structural and outcome factors of technology-mediated learning. Moreover, it considers the student learning process, which is crucial for the resulting learning outcomes and thus for the quality of technology-mediated learning (Gupta and Bostrom, 2009).

Technology-mediated learning is based upon two underlying assumptions. First, the framework assumes that a particular learning context influences the embedded learning structures (DeSanctis and Poole, 1994). In this dissertation, I follow a constructivist understanding of learning, describing that knowledge is always constructed by the learner itself and can be influenced by experiences (Von Glasersfeld, 1987). In this vein, the learning experience is influenced by a learning method. I am applying IT artifacts with an adaptive ability based on recent advantages in AI, in particular the classes of adaptive writing support systems and conversational agents, as a learning method structure. Second, the framework of technology-mediated learning focuses on the learning process of a learning activity. By definition, a learning process describes the sum of the single activities a student does. According to Gupta and Bostrom (2009), it describes how a student interacts with and adapts the learning method structures.

The process of a learner is a complex phenomenon and includes various variables such as cognitive processes and interactions related to the learning methods. By nature, it varies between individual learners, the received support in the learning process (e.g., based on adaptive IT), and other elements in a particular learning environment. Gupta and

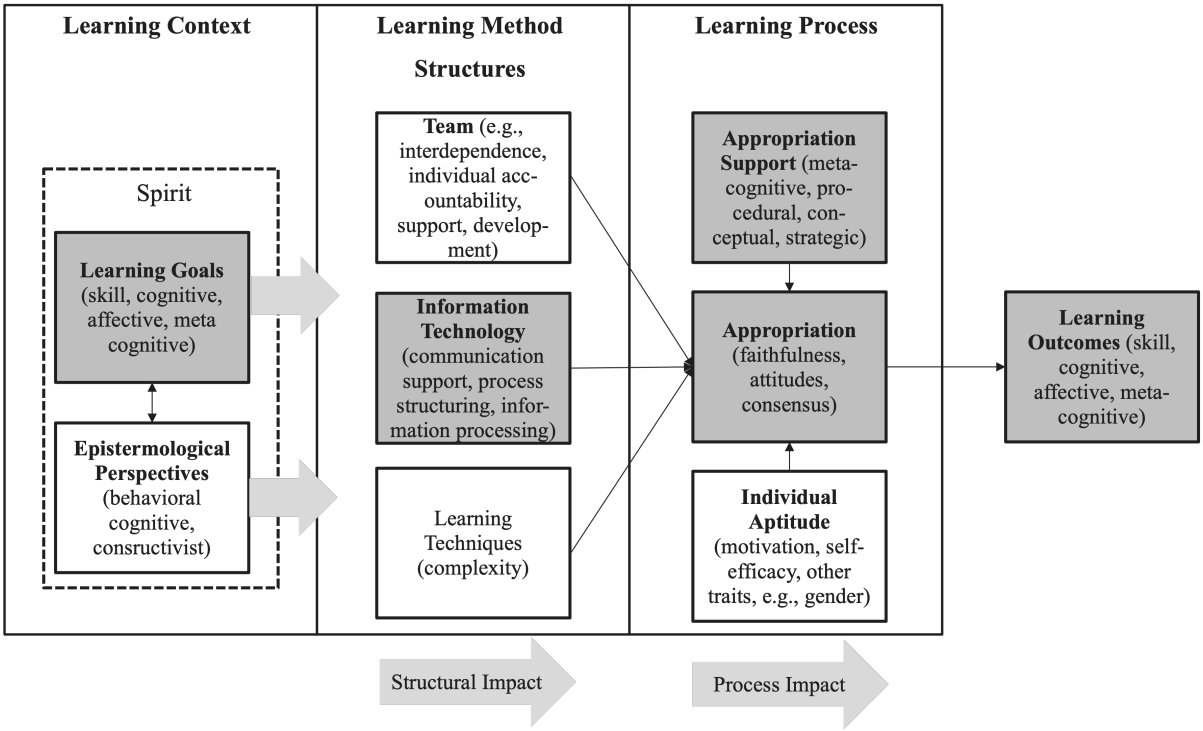



Figure 3: Focus of the dissertation based on the technology-mediated learning framework by Gupta and Bostrom (2009).
Source: Gupta and Bostrom (2009)

Bostrom (2009) define learning outcomes as the pivotal measurement of learning success as a result of a learning process in technology-mediated learning. Accordingly, learning outcomes are "the target assessment or measures used to determine the achievement of learning goals" (Gupta and Bostrom (2009), p. 691).

As described before, in this dissertation I follow a constructivist view of learning. The constructivist learning paradigm states that learning is achieved by students "constructing" knowledge from their own experiences (Von Glasersfeld, 1987). I aim to investigate in particular how novel adaptive writing support systems and pedagogical conversational agents, as a new class of IT, influence the student learning process and the resulting learning outcomes for argumentation skills. To emphasize the most important aspects, my objective is to focus on particular aspects of the technology-mediated learning framework of Gupta and Bostrom (2009). Hence, I focus on learning goals, IT, appropriation, and learning outcomes. My work puts rather little emphasis on the epistemological perspective, the team, the complexity of learning techniques, and individual student characteristics. The technology-mediated learning framework (Gupta and Bostrom, 2009) and the focus of my dissertation with regards to this framework are highlighted in Figure 3.

In the following, I will provide a more detailed background on the specific learning context of argumentation skill learning as well as on the particular learning method structures for

technology-mediated argumentation learning. In the different sections of my work, I will then refer to the dimensions of technology-mediated learning and highlight the differences in the learning process and the measured learning outcomes for the particular studies.

## 2.2 Learning Context: Argumentation Skill Learning

Technology-mediated learning distinguishes between two elements for the learning context: the learning goals and the epistemological perspectives Gupta and Bostrom (2009).

### 2.2.1 Learning Goals and Learning Outcomes

The learning goals comprise the knowledge expected to be learned after going through a certain learning process (Gupta and Bostrom, 2009). According to Anderson and Sosniak (1996), four different learning goals can be distinguished: skills, cognitive, affective, and meta-cognitive learning goals. The aim of technology-mediated learning is to achieve these learning goals and measure them as learning outcomes of the learning process (see Figure 3). For my work, I follow the definition of learning objectives based on the revised Bloom's taxonomy by Krathwohl (2002) (see Table 2). The taxonomy differentiates between four different dimensions of knowledge: *factual, conceptual, procedural, and cognitive* knowledge. Factual knowledge represents the knowledge of terminology and specific details (e.g., a historical fact or a certain date). Conceptual knowledge encompasses the knowledge of theories and models. It relates to the interdependence among basic concepts and the metacognitive knowledge of a learner. The procedural knowledge dimensions refer to subject-specific skills and methods. It refers to process-related knowledge such as knowing how to do a certain task in the right way (e.g., knowing how to orchestrate and structure a project to achieve something). Metacognitive knowledge is about strategic and cognitive knowledge (such as argumentation, communication, or problem-solving).

| | **remember** | **understand** | **apply** | **analyze** | **evaluate** | **create** |
|---|---|---|---|---|---|---|
| **Factual** | | | | | | |
| **Conceptual** | | | | | | |
| **Procedural** | | | | | | |
| **Metacognitive** | | | | | | |

Table 2: Revised taxonomy of learning outcomes according to Krathwohl (2002).
Source: Krathwohl (2002)

Krathwohl (2002) also distinguishes between the cognitive processes of various learning tasks. The categories are *remember*, *understand, apply, analyze, evaluate*, and *create* (Krathwohl, 2002). These six categories are considered to be hierarchically ordered in their

complexity, meaning creating something is more complex and usually involves evaluating and analyzing something. Learning and training "higher" learning goals is usually one of the objectives of higher education (Krathwohl, 2002). The more the learning outcomes are on the level of "evaluate/create", the more complex the learning process for the students.

In this dissertation, my objective is to help students to learn how to argue based on recent methods in NLP and ML. Hence, my learning objective is to support metacognition skill learning. I do that by supporting students in various learning tasks, where they mostly have to analyze, evaluate, and also create knowledge.

To better understand the concept of metacognitive argumentation skill learning, I aim to provide the reader a brief background on the concept of argumentation, on the underlying argumentation models, and on argumentation skill learning.

### 2.2.2 The Concept of Argumentation

As Kuhn 1992 states, the skill to argue is of great significance not only for professional purposes such as for collaboration and for solving difficult problems but also for most of our daily life. Argumentation is an omnipresent rudiment of daily communication and thinking (Kuhn, 1992). In general, argumentation aims to increase or decrease a controversial standpoint's acceptability (Eemeren et al., 1996). Logically, structured arguments are a required precondition for persuasive conversations, general decision making, and drawing acknowledged conclusions (Duschl and Osborne, 2002). Across multiple fields, including law, science, politics, or management, individuals must support their claims with essential facts, argue in support of conclusions derived from those facts, and counter their opponent's claims in a principled way to convince others of their position or justify a conclusion (von Aufschnaiter et al., 2008). Over several decades, research has demonstrated that humans are generally deficient in argumentation (e.g., Byrne (1989); Marcus and Rips (1979); Tversky and Kahneman (1974)): they often fail to recognize the difference between merely expressing an opinion and making a fact-based claim. Moreover, they do not rebut others' arguments but ignore points of conflict and persist with their own arguments (Byrne, 1989). Thus, research in fields such as human—computer interaction or IS has demonstrated increased interest in developing argumentation systems to support individuals (Eemeren et al., 1996).

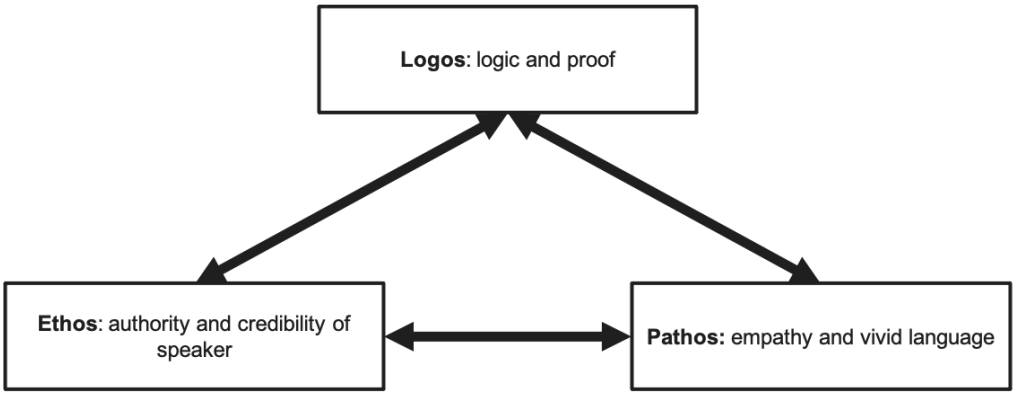



Figure 4: Aristotle's theory of persuasion based on logos, ethos, and pathos.
Source: Own illustration

### 2.2.3 Argumentation Theory

Argumentation theories have a long history in philosophy, linguistic research, and mathematics. Aristotle provided one of the first foundations in his theory of persuasion. Many frameworks and "rules" of argumentation have since been proposed and identified (e.g., Kuhn (1992); Toulmin (1984); Walton (1996); Walton et al. (2008); Toulmin (2003); Eemeren et al. (1996)). In his fundamental theory, Aristotle distinguished between three interconnected principles of persuasion (see Figure 4)—logos (logic and proof of argumentation), ethos (authority and credibility of the speaker), and pathos (empathy and vivid language)—on which an individual can build in their effort to persuade an opponent. Logos focuses on the general formality and structure of argumentation; ethos depends largely on the individual and their relationship to the opponent, and pathos concerns context-related emotions and the strength of languages.

Formal logic, a branch of mathematics, has long been the main perspective on persuasive argumentation (e.g., North Whitehead and Russell (1910); Scheuer et al. (2010)). More recently, however, theories have also focused on practical human argumentation (e.g., Kuhn (1992); Toulmin (2003)). Most theoretical and practical approaches to argumentation vary in their level of detail, perspective, and specific context of applicability. Nevertheless, several scholars (i.e., Scheuer et al. (2010, 2012a); Scheuer (2015)) have observed that logic is likely the foundation that underlies different theoretical argumentation approaches. Argumentation theory generally agrees on the importance of logos as a basis for proper argumentation. Argumentation discourse should consider all relevant facts, claims should be well-grounded and supported by premises, and both supporting and conflicting claims should be taken into account (Kuhn, 1992; Toulmin, 2003).

In most of the past argumentation learning research, scientists only focused on the "logos" level of argumentation and persuasion. Also, in this research I mostly investigate the

impact of logical argumentation support (logos level) on students' formal argumentation skills (e.g., in section 5 or section 7).

### 2.2.4 Argumentation Models

Most argumentation theories formulate formal argumentation models that address the logical component of Aristotle's theory of persuasion. The Toulmin model is among the most prominent of these (Toulmin, 1984, 2003). Accordingly, an argument consists of several components, including a claim and at least one premise. The claim is the central component and statement which is justified by premises. According to Toulmin's argumentation theory, a good argument involves a logical structure founded on grounds, claim, and warrant, whereas the grounds are the evidence used to prove a claim (Toulmin, 1984). For example, according to the Toulmin Model, each argument can be broken down into six parts (claim, premise/ ground, warrant, backing, rebuttal, qualifier). Each component fulfills a specific argumentative role and complements the others to create the argument. Figure 5 illustrates the six elementary components according to the Toulmin Model to which each argument can be broken down.

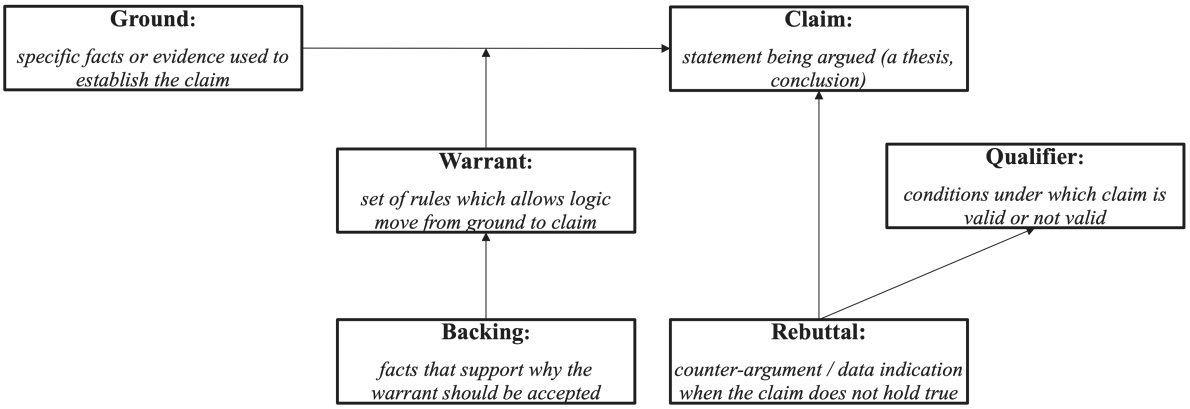



Figure 5: The six elements an argument can be broken down according to the Toulmin model.
Source: Toulmin (2003)

Each of these argument components fulfills a specific argumentative role and complements each other to create the argument:

1. **Claim**: The central point or conclusion of the argument. The ultimate goal is to convince the opponent of the truthfulness of the claim.
2. **Ground** (Fact, Evidence, Data): The support or rationale for the claim. It is the data or evidence that the arguer uses to explain and support their claim.

3. **Warrant**: The inference connection between the claim and ground.

4. **Backing**: Similar to the ground, the backing provides support to the warrant. It is often used when the warrant itself is not convincing enough.

5. **Rebuttal**: Potential objection to the given claim. According to Toulmin, it is important that the arguer shows that they are aware of possible opposition to the claim.

6. **Qualifier**: Indicator of the degree of certainty associated with the claim. It allows the arguer to add nuance to the claim.

Nevertheless, across multiple argumentation models, the claim and premise (also called evidence) are generally considered the main components of every argument (Toulmin, 1984, 2003; Walton et al., 2008; Walton, 1996). The other components are considered supporting sub-argument parts that may or may not exist in an argument. Hence, in this dissertation, I mainly follow the simplified version of a claim - premise argumentation model as done in most literature about argumentation skill learning and argumentation mining (Figure 6, Scheuer (2015); Stab and Gurevych (2014a, 2017); Lawrence and Reed (2019)):

Figure 6: Basic argumentation model with the logos components claim and premise. Source: Own illustration

### 2.2.5 Argumentation Skill Learning

Nevertheless, dialectical concepts for teaching argumentation are scarce. Jonassen and Kim (2010), p. 442 identified three major causes for that: "*teachers lack the pedagogical skills to foster argumentation in the classroom, so there exists a lack of opportunities to practice argumentation; external pressures to cover material leaving no time for skill development; and deficient prior knowledge on the part of learners*". Therefore, many authors have claimed that fostering argumentation skills should be assigned a more central role in our formal educational system (Driver et al., 2000; Kuhn, 2005). Most students learn to argue in the course of their studies simply through interactions with their classmates or teachers. In fact, individual support of argumentation learning is missing in

most learning scenarios. However, to train skills such as argumentation, it is of great importance, for the individual student to receive continuous feedback, also called formative feedback, throughout their learning journey (Hattie and Timperley, 2007). According to Sadler (1989), the outcome of feedback is a specific information relating to the task or process of learning that fills a gap between what is understood and what is aimed to be understood. Even in fields where argumentation is part of the curriculum, such as law and logic, a teacher's ability to provide feedback is naturally limited by constraints on time and availability. Especially in more common large-scale lectures, the ability to support a student's argumentation skills individually is hindered, since for teachers and professors, it is becoming increasingly difficult to provide ongoing and individual feedback to a single student.

### 2.2.6 Constructivist Learning Context

In order to teach argumentation skills, the epistemological perspective is an important factor to design a certain pedagogical scenario and thus the learning method structures. In this dissertation, I do not aim to focus on the epistemological component of learning. However, I will briefly explain the underlying constructivist learning context, which I apply in this work.

The epistemological perspectives distinguish between different paradigms for the overarching beliefs about the nature of knowledge and what it means to gather knowledge (Leidner and Jarvenpaa, 1995). The most common classification in education distinguishes between *behaviorist, cognitivist*, and *constructivist* views of learning (Leidner and Jarvenpaa, 1995). While behaviorism assumes that learning occurs when students exhibit a new behavior or change their behavior in response to instructional stimulations (e.g., a classical presentation of a teacher), the cognitivist perspective is based on the assumption that the cognitive structures in a learner's brain is most significant to the learning process as they process information (Gupta and Bostrom, 2009).

The constructivist learning view regards learning as a change in meaning resulting from experience (Von Glasersfeld, 1987). Accordingly, knowledge (and also the truth) is constructed by learners themselves (Gupta and Bostrom, 2009). The behaviorist and cognitivist paradigms share the common assumption that the learning of a student must be externally induced by behavior (Ertmer and Newby, 1993). In contrast, constructivist learning argues that students construct understanding. A learner does not only mirror what they perceive, they search for comprehension in what they experience in the world, even if they do not have complete information (Von Glasersfeld, 1987). Hence, the idea of constructivist education research comprises that students actively construct their own

knowledge. Accordingly, learning itself is described as an "active mental work, not passive reception of teaching" (Woolfolk Hoy, 2000).

Based on a constructivist perspective, in this work, I mainly apply and contribute to two common learning theories, which are based on a constructive learning paradigm: the interactive, constructive, active, and passive (ICAP) framework by Chi and Wylie (2014), and self-regulated learning theory (Bandura, 1991; Zimmerman and Schunk, 2001). In the course of my dissertation, I will link these theories to the content and explain how I believe these theories might support my underlying hypothesis that adaptive argumentation skill learning based on individual learner-centered argumentation tools helps students to construct new knowledge. To avoid repetitions, I will briefly explain the basics of both theories in the following paragraphs.

#### 2.2.6.1 ICAP Framework

According to the ICAP Framework (i.e., interactive, constructive, active and passive Framework) by Chi and Wylie (2014), the learners' engagement with learning materials can range *"from passive to active to constructive to interactive"* (Chi and Wylie, 2014) and will result in an improved learning outcome. Whereas in passive engagement students only consume or receive learning materials, in active engagement students actively manage the content presentation. In the two most engaging forms of interaction according to Chi and Wylie (2014), students deepen their interaction, e.g., by comparing the learning materials with prior knowledge (constructive engagement), by debating with others or asking and answering questions (interactive engagement). Each mode of the ICAP framework corresponds to different types of behaviors and knowledge change processes predicting different learning outcomes (Chi and Wylie, 2014).

Table 3 provides an overview of the four interaction modes of students and outlines three examples of how the interaction mode affects the students' learning activities. Passive student behavior describes the activity of students engaging with the instructional information presented without additional physical activity, e.g., only reading a certain text. Active student behavior comprises a physical activity of the student, e.g., highlighting important text passages. Constructive student behavior is defined by students actively constructing new information such as ideas that are beyond the presented material. An example would be to sketch a mind map of the highlighted text passages. Interactive student behavior requires students to actively engage on a certain topic, for example, by debating about the pros and cons of a certain position. This can be done by "dialoguing extensively on the same topic, and not ignoring a partner's contribution" (Chi and Wylie,

| | **Passive** *Receiving* | **Active** *Manipulating* | **Constructive** *Generating* | **Interactive** *Dialoguing* |
|---|---|---|---|---|
| **Listening** to a lecture | Listening without doing anything else but oriented toward instruction | Repeating or rehearsing; copying solution steps; taking notes | Reflecting out loud; drawing concept maps; asking questions | Defending and arguing a position in small groups |
| **Reading** a text | Reading entire text passages silently/ aloud without doing anything else | Underlining or highlighting; summarizing by copy and delete | Self-explaining; integrating across texts; taking notes in one's own words | Asking and answering comprehension questions with a partner |
| **Observing** a video | Watching the video without doing anything else | Manipulating the tape by pausing, playing, fast-forward, rewinding | Explaining concepts in the video; comparing and contrasting to prior knowledge or other materials | Debating with a peer about the justifications; discussing similarities and differences |

Table 3: Exemplary learning activities by mode of engagement according to the ICAP Framework from Chi and Wylie (2014).
Source: Chi and Wylie (2014)

2014). The more interactive the mode of engagement is, the more effective the learning activity (Chi and Wylie, 2014).

Following this hypothesis, adaptive and interactive IT-based learning systems are capable of fostering the students' engagement, as they add the new component of interactivity or even dialoguing to argumentation learning scenarios. Compared to common argumentation learning systems, pedagogical conversational agents, for example, are able to discuss and tutor students through the learning content, just like human instructors would. In fact, the successful application of pedagogical conversational agents to adaptively meet the individual needs of learners and to increase their learning outcomes has been demonstrated for learning various skills, such as problem-solving skills (Winkler et al., 2020), programming skills (Hobert, 2019), mathematical skills (Cai et al., 2019) as well as for learning factual knowledge (Ruan et al., 2019), but it has not been investigated for argumentation skills.

Apart from the explained ICAP framework, self-regulated learning theory regards the combination of a learner's individual goal, motivation, and feedback through self-monitoring and self-evaluation during the learning process as beneficial for learning outcomes (Bandura, 1977, 1991; Zimmerman and Schunk, 2001; Bjork et al., 2013).

#### 2.2.6.2 Self-regulated Learning

Self-regulated learning theory supports the underlying hypothesis that the combination of a goal setting structure with individual self-evaluation and self-monitoring (e.g., in the form of formative feedback Bandura (1977, 1991)) on a student's learning outcome increases the student's self-efficacy (Zimmerman and Schunk, 2001; Zimmerman, 1989; Bjork et al., 2013). Self-regulated learning theory reflects that students learn better with formative feedback and goal setting (Bandura, 1991). Especially for students in a learning process, critical reflection through self-monitoring and self-evaluation is an important component for effective learning (Zimmerman and Schunk, 2001), also reflected in literature on transformation learning (e.g., Mezirow (1991)). It can be an initial trigger for a student's learning process and thus the creation of new knowledge structures (Piaget et al., 1986). However, the right portion of self-monitoring and self-reflection in combination with a learning goal is important for students to learn effectively (Bandura, 1991; Zimmerman and Schunk, 2001). Therefore, in this dissertation, I hypothesize that the right level of feedback on a students' skill, such as argumentation or empathy, could lead to more self-efficacy and thus to motivation to learn how to react to other people's perspectives in an appropriate manner. However, I do not aim to focus solely on self-regulated learning and the learner's entire cognitive process but rather on the proven self-reliant learning so that individuals can learn argumentation skills irrespective of an instructor, time, and location. Being aware of the fact that self-regulated learning involves more than formative feedback, self-monitoring, and self-evaluation, I still focus on these (individual) concepts — in combination with other guiding literature — with the simple objective of designing novel argumentation learning tools to better support humans in learning argumentation skills based on recent technological advances embedded in known theoretical concepts (such as self-regulated learning).

More recent aspects that are relevant for self-regulated learning theory focus on learning from errors through self-regulated learning. According to Bjork et al. (2013), accurate self-monitoring and evaluation of one's learning progress are fundamental components for effective self-regulation of learning. Nevertheless, humans struggle to monitor and evaluate their learning and comprehension of complex materials (Bjork et al., 2013). Hence, the design of appropriate monitoring and evaluation elements for certain skills might help learners to continuously learn more effectively from a long-term perspective (i.e., Roediger and Karpicke (2006); Zimmerman and Schunk (2001)). One way to facilitate effective monitoring and an evolution of learning is to combine skill monitoring with formative assessment and performance feedback (Black and Wiliam, 2009; Hattie and Timperley, 2007). To be effective, feedback should define goals, monitor the progress toward the goals, and identify the activities that will bring the learner toward the goals (Hattie and

Timperley, 2007). Furthermore, research has suggested that feedback on learners' errors can greatly facilitate new learning (Metcalfe, 2017). According to (Metcalfe, 2017), p. 484, making mistakes and receiving feedback "enhances later memory for and generation of the correct response, facilitates active learning, [and] stimulates the learning to direct attention appropriately."

The described findings on "learning from errors" (e.g., Bjork et al. (2013)) and formative feedback (e.g., Hattie and Timperley (2007)) play a significant role when designing AI-based learning systems. Hence, I will refer to them again in different sections of my dissertation, such as when designing an AI-based argumentation writing support system for peer reviews in section 5.

## 2.3 Learning Method Structures: Technology-Mediated Argumentation Skill Learning

Technology-mediated learning learning not only considers the learning context, such as the underlying educational concepts and the learning goals, but also the learning method structures, which aim to structure and facilitate the learning process and thus the learning outcomes of students. After explaining the concept of argumentation skill learning and the background of constructivist learning, the next paragraphs will include a short overview of technology-mediated learning tools for argumentation, especially of the classes of computer-supported collaborative learning and intelligent tutoring systems. In this vein, I will take a deep dive into the five different classes of argumentation learning systems: representational guidance approaches, discussion scripting approaches, adaptive support approaches, AI-based argumentation writing support systems, and conversational argumentation learning systems.

Different technological approaches have been used for learning method structures. Research started by using film, sound, radio, or TV in the classroom (Scheuer, 2015). The appearance of personal computer technology intensified the research on educational technology and technology-mediated learning (Koschmann, 1996). Early classifications of computer-based learning, such as of Koschmann (1996), distinguished between four main paradigms in the field of educational technology. Each paradigm has specific assumptions about learning and teaching and utilizes specific technological and research approaches. Koschmann (1996) distinguished between computer-assisted instruction, intelligent tutoring systems, Logo-as-Latin, and computer-supported collaborative learning. The paradigms of computer-supported collaborative learning (Dillenbourg et al., 2009; Koschmann, 1996) and intelligent tutoring systems are particularly relevant to argumen-

tation learning (Scheuer, 2015), since argumentative discussions and debates have been identified as key factors for collaborative learning settings in particular. In the following sections, I will shortly introduce these two classes more specifically to the reader.

### 2.3.1 Computer-Supported Collaborative Learning

Computer-supported collaborative learning emerged as a research area for educational technology in the 1990s (Scheuer et al., 2010; Dillenbourg and Hong, 2008). The rapid technological advances because of the use of computers at home and the connectivity through the world wide web fueled the research community to investigate how to leverage these technologies to improve educational scenarios. Moreover, scholars also generally focused more on the potential of collaborative learning approaches, especially for higher learning goals (Scheuer, 2015). Hence, the underlying concepts of computer-supported collaborative learning focus on the interaction between learners as the primary source of learning (Dillenbourg and Hong, 2008). The majority of computer-supported collaborative learning research interventions aim to investigate how technology can facilitate and scaffold interactions between learners to achieve certain learning outcomes.

Technology-mediated learning tools have been designed to facilitate student peer feedback (Rietsche and Söllner, 2019), to support student peer assessments in large-scale learning environments (Lehmann et al., 2016), to better structure argumentative discussions between learners (Fischer et al., 2013), or to support empathy self-monitoring in student-based discussions (Santos et al., 2018). Computer-supported collaborative learning learning tools have been proven to facilitate, scaffold, and support especially higher learning goals in large-scale learning scenarios (e.g., Noetel et al. (2021); Dillenbourg et al. (2009); Dillenbourg and Hong (2008); Fischer et al. (2013); Santos et al. (2018)).

### 2.3.2 Intelligent Tutoring Systems

Intelligent tutoring systems research traditionally focuses more on the individual learner and how technology can be used to intelligently tutor a student through a specific learning process. Research on intelligent tutoring systems has a long history and dates back to the 1970s (e.g., Atkinson and Shiffrin (1968); Suppes and Morningstar (1969)). Scholars investigating intelligent tutoring systems traditionally focused on the learner's mental representations of knowledge and therefore on the cognitive components. Different techniques, especially from early AI, are utilized to model mental structures, skill levels, and learning processes of students to improve the computer-based interaction with the individual (Corbett and Anderson, 1994). The term cognitive tutors emphasized the

ability of technology-mediated learning to model the cognitive structures of students to provide them with feedback on their particular learning process (Anderson et al., 1995). Early research studies showed that students were gaining knowledge most efficiently when computer-based tutors were providing ongoing and rapid feedback, consisting of short and directed error messages. While in the 1990s, research found that computer-based tutors are able to assist students more effectively if they present themselves as nonhuman tools rather than as a virtual human educator (Anderson et al., 1995), recent research showed significant effects of anthromorphization of pedagogical agents in different contexts (i.e., Noetel et al. (2021)).

Common methods to model students' learning processes are, for example, *model tracing*, *buggy rules*, or *knowledge tracking* (Corbett and Anderson, 1994). Whereas model tracing is a computational method to model the mental operation of a students' problem-solving process by analyzing different user inputs, knowledge tracking is used to model the probabilities that a learner has achieved a certain knowledge module or skill based on the analysis of his or her individual learning process steps (Corbett and Anderson, 1994; Anderson et al., 1995). The different approaches are used to then provide the learner with a learning experience with an artificial tutor similar to a real educator. A student receives intelligent recommendations (Bauman and Tuzhilin, 2018), scaffolded learning tasks (Winkler et al., 2020; Janson et al., 2019), or individual feedback based on his or her knowledge, skill level, or individual errors (Ruan et al., 2019).

With the rapid technological advances in NLP, ML, and AI as well as the widespread storage of learner data in learning management systems, the modeling of cognitive processes, knowledge levels, or learning processes has received continuous attention in different disciplines. Leveraging the approach of intelligent tutoring systems for digital education comes with certain advantages that are worth mentioning compared to other educational technology approaches. First, intelligent tutoring systems are able to adjust to a certain learner's skill profile and provide tailored recommendations, feedback, or scaffolds based on the learner's misconceptions. With the wide-scale availability of digital learning traces from common learning management systems, learning profiles, or user-interaction data, intelligent tutoring systems are promised to provide a personalized and unique learning experience similar as a human educator would provide but independent of time, availability, and location of a human instructor or peers. Through intelligent tutoring systems, students are promised to learn where they want, how they want, and whenever they want and receive the most tailored education according to their learning profile (Noetel et al., 2021; Xu et al., 2021). With the fine-grain analysis of learning interaction data and learning profile, intelligent tutoring systems are able to model and facilitate a learning process on various granularity levels. Especially deep learning and transfer learning approaches promise to improve these models significantly Rosé et al. (2008).

### 2.3.3 Pedagogical Conversational Agents

A subset of intelligent tutoring systems that has recently received attention are pedagogical conversational agents. Conversational agents are computer programs that are designed to communicate with users through natural language interaction interfaces (Rubin et al., 2010; Shawar and Atwell, 2005). In today's world, conversational agents, such as Amazon's Alexa, Google's Assistant, and Apple's Siri, are ubiquitous, with their popularity steadily growing over the past few years (eMarketer, 2017; Krassmann et al., 2018). They are implemented in various areas, such as customer service (Xu et al., 2017), healthcare (Kowatsch et al., 2017; Laumer et al., 2019), or education (Kerly et al., 2007; Winkler and Söllner, 2018). Hobert and Wolff (2019) define conversational agents in the education domain as distinct forms of learning applications that provide an individual and personalized interaction with students. There is growing consensus that the interaction with conversational interfaces benefits students in their learning process. First of all, conversational agents offer a natural and intuitive way for students to express themselves in computer-mediated educational settings, which in turn drives learning outcomes (e.g., Wambsganss et al. (2021c); Ruan et al. (2019); Winkler et al. (2020)). Second, conversational agents allow for an adaptive interaction path that goes beyond a static user interaction experience, thereby increasing student engagement (Winkler et al., 2021). Third, the design of conversational agents often incorporates anthropomorphic features that further support educators in building trustworthy and engaging learning scenarios (Winkler and Söllner, 2018; Noetel et al., 2021).

Inspired by the potential advantages of conversational over non-conversational user interfaces, the development of conversational agents in education is based on the design of intelligent tutoring systems (e.g., Atkinson and Shiffrin (1968); Suppes and Morningstar (1969)). Similar to a human tutor, intelligent tutoring systems and especially conversational systems can present instructions, ask questions, and provide immediate feedback (Kulik and Fletcher, 2016). Commonly referred to as pedagogical conversational agents, these pedagogical systems evolved from abstract entities with limited technological possibilities to systems that are able to interact with learners using multiple channels of communication, exhibit social skills, and perform different roles, such as tutors (Payr, 2003; Wambsganss et al., 2021c), motivators, learning companions (Kim and Baylor, 2008; Winkler et al., 2020), or learning support (Song et al., 2017; Hobert and Wolff, 2019; Winkler and Söllner, 2018).

However, the development of pedagogical conversational agents and intelligent tutoring systems is a complex matter because of the interdisciplinary complexity of modeling the learning process based on statistical approaches. Moreover, the learner–computer inter-

action often plays the most crucial role in the design and development of technology-mediated learning tools that are able to impact students' learning outcomes significantly (Noetel et al., 2021; Xu et al., 2021).

### 2.3.4 Argumentation Learning Systems

Based on argumentation theory and, most prominently, the Toulmin model, researchers have developed different tools to support and teach argumentation: over a period of 35 years, more than 60 tools have been developed and published to support learners in creating, editing, interpreting, or reviewing arguments (e.g., Afrin et al. (2021); Chernodub et al. (2019); Huang et al. (2016); Pinkwart et al. (2009); Smolensky et al. (1988); Suthers and Hundhausen (2001); Wang et al. (2020)).

Especially, argumentation has emerged as a focus area in computer-supported collaborative learning. Argumentation research on intelligent tutoring systems centers more on analyzing, modeling, and supporting technology-enhanced argumentation learning activities in specific domains (Dillenbourg and Hong, 2008; Nye, 2014; Scheuer, 2015). The combination of computer-supported collaborative learning and intelligent tutoring systems to support collaboration and argumentation both adaptively and individually is a relatively new research area (Fischer et al., 2013). Researchers have designed and evaluated several tools based on input masks and representational guidelines to support high school students' active writing processes. This has been investigated across various fields, including law (Pinkwart et al., 2009), science (Osborne et al., 2016; Suthers and Hundhausen, 2001), and conversational argumentation (De Groot et al., 2007). Based on the classification of Scheuer (2015), three different IT-based argumentation learning systems can be distinguished: representational guidance approaches, discussion scripting approaches, and adaptive support approaches. I will briefly outline each of these different approaches and present examples of past argumentation learning systems. With this dissertation, I introduce the design and concept of two novel approaches to adaptive argumentation skill learning based on AI: AI-based argumentation writing support systems and conversational argumentation learning systems. The distinct characteristics of the approaches and their differences are summarized in Table 4.

#### 2.3.4.1 Representational Guidance Approaches

Representational guidance approaches are based on the idea that knowledge can be represented in different formats. The most common representation of knowledge is probably

in textual form (e.g., textbooks). Other representation formats can include diagrams, concept maps, or visualizations to make certain information more accessible to the reader (Scheuer et al., 2010; Suthers and Hundhausen, 2001).

These different knowledge representations are utilized and manipulated to influence students' learning processes. Technology-mediated representational tools can support students' learning with different knowledge representations by facilitating knowledge exploration, creation, modification, or organization (Scheuer, 2015).

Research has, therefore, focused on if and how different knowledge representations of argumentation structures can support argumentation skill learning. Figure 7 displays the architecture of the representational guidance system of Suthers (2003). Suthers (2003) supported students' argumentation learning by providing representations of argumentation structures with the objective of stimulating and improving individual reasoning, collaboration, and learning. Their approach represents argumentation discourse in the form of linked graphs. One instantiation of their architecture was focusing on the representation of displaying argumentation as *text-based notations, graph-based notations, and table-based notations.*. Figure 8 provides an overview of the user interface of their system Belvedere (Suthers and Hundhausen, 2001; Suthers, 2003).

The system class of representational guidance approaches formed a popular research subject especially in the early 2000s. Having the objective to stimulate and support argumentation competencies of students, several authors investigated how different representations can support argumentation skill learning (e.g., Nussbaum et al. (2007); Suthers (2003); Pinkwart et al. (2009); Bell (1997)). Nevertheless, these approaches mostly focus on the representation of argumentation, not on the learning itself. Moreover, one of the limitations of theses approaches come with the technical foundations, i.e., that these presentations are usually built on domain-dependent ontologies and are thus not suitable for a wide-scale adoption (Scheuer et al., 2010). Moreover, larger argumentation graphs tend to confuse users and are not useful for longer writing tasks.

As Scheuer (2015) noted, it remains open "how the different notations impact and influence learning processes and learning outcomes" (Scheuer (2015), p. 106). Suthers (2003) addressed some of these questions with his representational guidance theory. For example, their observations indicate that graphs and tables can represent argumentative relations in a more vivid way compared to written text (Suthers, 2003). In conclusion, learners are more likely to elaborate on such argumentative relations when creating or analyzing arguments.

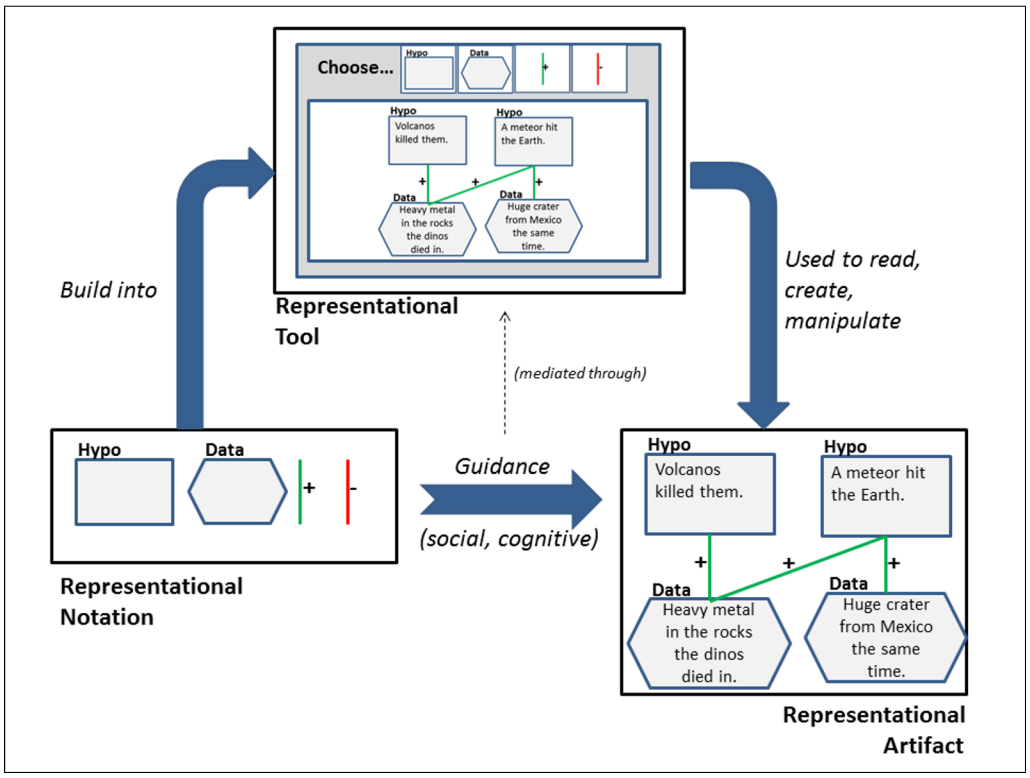



Figure 7: Basic concept of the representational guidance approach of Suthers (2003); different representations of argumentation structures are manipulated to guide the argumentation skill learning of students.
Source: Scheuer (2015); Suthers (2003)

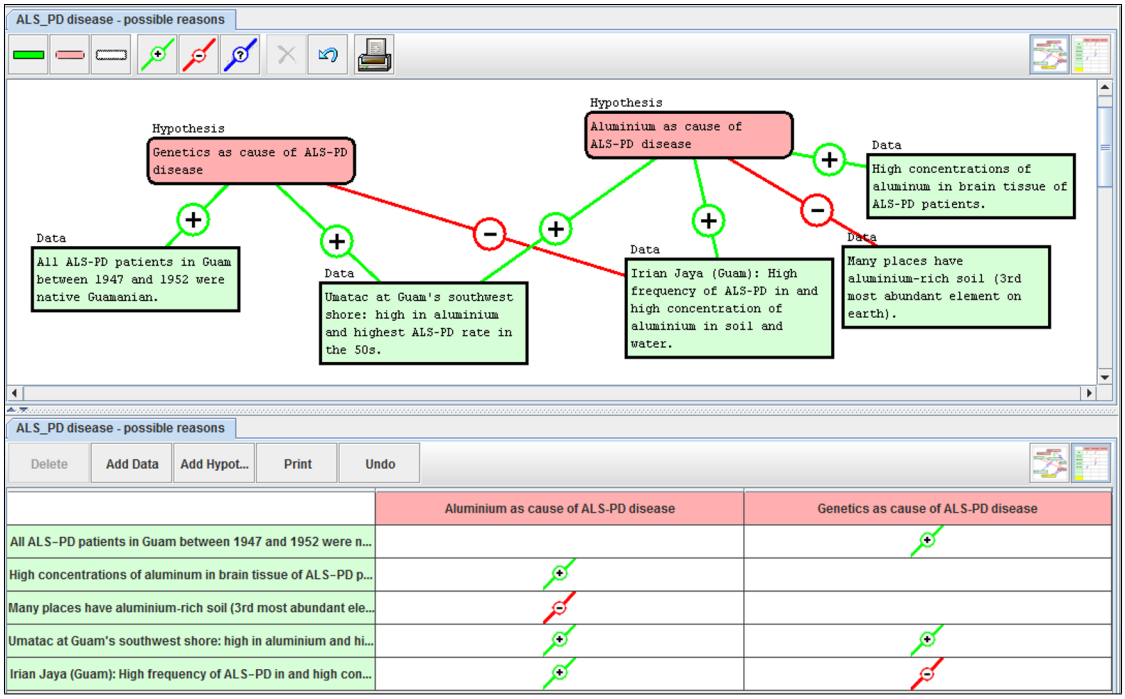



Figure 8: Screenshot of the Belvedere system (Suthers et al. 2001). It uses a node-and-link graph and a table to express hypotheses-evidence relationships.
Source: Suthers (2003); Scheuer (2015)

#### 2.3.4.2 Discussion Scripting Approaches

Based on the constructivist learning paradigm, discussion scripting approaches aim to provide learners with external facilitation and structures in forming arguments for discussing and communicating with others. When interacting in collaborative learning exercises, sometimes students are highly motivated. However, they lack a certain pedagogical underlying concept. Hence, scripting approaches aim to provide a scaffold for the learner to guide discussions and ultimately support argumentation learning.

Different concepts have been investigated by various scholars for collaboration scripts in both face-to-face and technology-mediated collaborative learning (e.g., O'Donnell et al. (2002); Kollar et al. (2006); Dillenbourg (2002b)). For example, Kollar et al. (2006) defined collaboration scripts as "scaffolds that aim to improve collaboration through structuring the interactive processes between two or more learning partners." (Kollar et al. (2006), p. 1). In their approach they identified five components for these scripts on how students should be guided in collaborative learning exercises: *learning objectives, type of activities, sequencing, role distribution*, and *type of representation*.

Dillenbourg (2002b) describes different kinds of collaboration scripts, including *induced scripts* (conveying expectations regarding desired patterns of interaction), *instructed scripts*

(instructions prior to collaboration), *trained scripts* (prior training of students regarding expected behaviors), *prompted scripts* (introduction to a certain behavior), and *follow-me scripts* (enforcing certain forms of interaction through a learning tool). According to Dillenbourg (2002b), the appropriate design of scaffolded scripts is, hence, very important for students' learning processes, their motivation, and their learning outcomes.

Building on prior approaches, Fischer et al. (2013) proposed a theoretical foundation for discussing scripting approaches by deriving their *theory of guidance*. Their underlying hypothesis assumes that collaborative learning is facilitated by mental knowledge structures, which they call internal collaboration scripts. Fischer et al. (2013) defined different components of these internal scripts (play, scenes, roles, and scriptlets) and a set of principles for each of these components, e.g., the *script configuration principle* (Fischer et al., 2013).

Based on these principles, researchers investigated different learning tools to improve the quality of argumentation through structured communication interfaces. The interventions were mostly on the scriptlet's level, also referred to as micro scripts, with the aim to guide a certain pedagogical concept (Dillenbourg and Hong, 2008; Fischer et al., 2013).

A popular approach is exemplarily illustrated by the tool of Weinberger et al. (2010) (see Figure 9), which was designed to scaffold the creation of individual messages through a form-like interface in an online discussion. This tool was used in several studies, for example, in Stegmann et al. (2012); Fischer et al. (2013); Weinberger et al. (2010). Students are provided with text fields to enter a *claim*, *grounds*, and *qualifications*. The argumentation structure is based on the Toulmin model of argumentation (Toulmin, 1984). Several empirical studies indicated that this pre-structuring of argumentation improves the formal quality of argumentation (e.g., Stegmann et al. (2012); Fischer et al. (2013)).

Another very widespread approach is the use of sentence openers to script argumentative discussions (see exemplary system AcademicTalk by McAlister et al. (2004) in Figure 10). These sentence openers usually consist of predefined phrases with keyword structures where students can choose to start the answer to a certain argument or engage in a discussion. A number of these argumentation learning systems exist, and they were mostly embedded in chat and threaded discussions to enhance the argumentation discussion quality, for instance, *Group Leader Tutor* by Mcmanus and Aiken (1996), the system *AcademicTalk* by McAlister et al. (2004), or *InterLoc* by Ravenscroft (2007).

Nevertheless, discussion scripting approaches come with several limitations on their adaption and learner-centered availability in the field. As Scheuer (2015) notes, discussion scripting approaches often lack adaptivity in scripted learning. Pre-defined sentence openers or form-like interfaces miss individual support for students in their personal learning

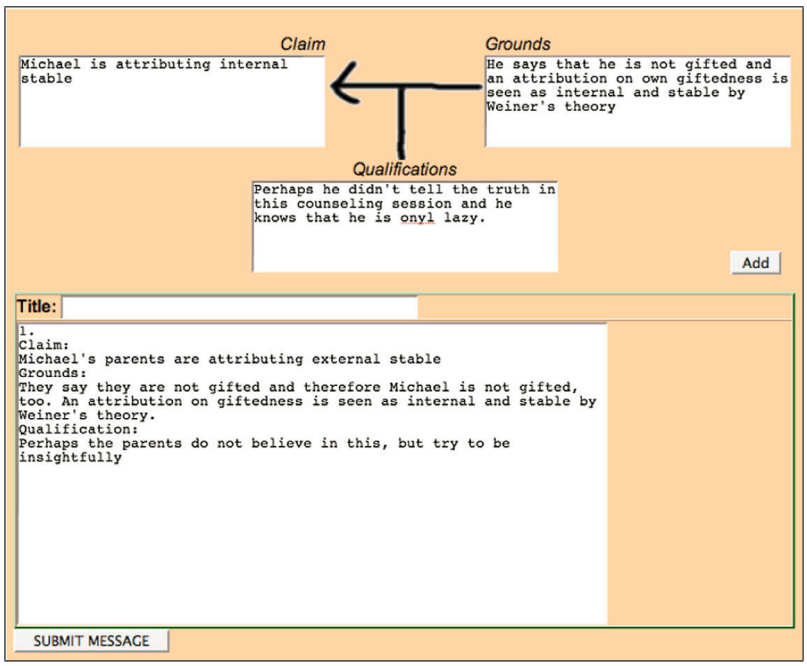



Figure 9: Learning tool by Weinberger et al. (2010) to support the creation of argumentative messages through a form-like interface in a forum discussion.
Source: Weinberger et al. (2010); Scheuer (2015)

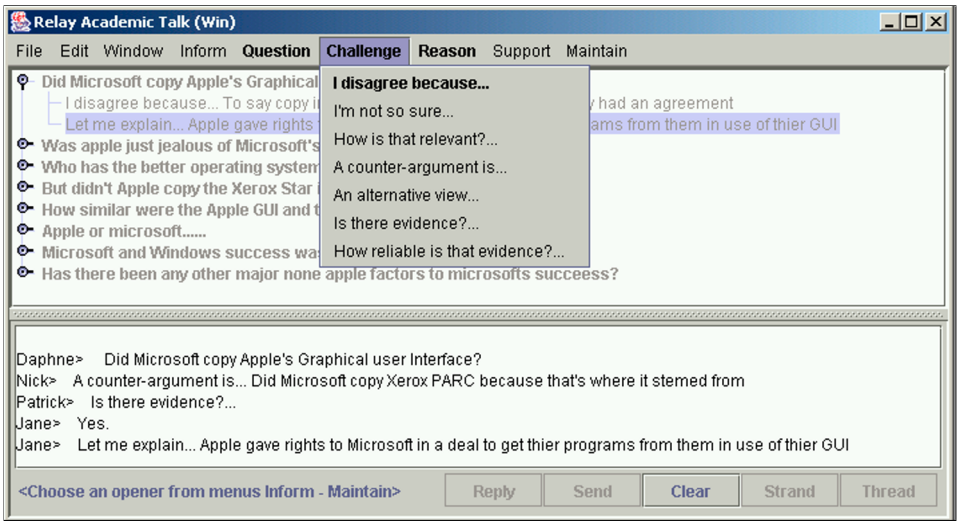



Figure 10: AcademicTalk system by McAlister et al. (2004): Students are provided with sentence openers to facilitate a discussion.
Source: McAlister et al. (2004); Scheuer (2015)

experience. Moreover, these systems are usually designed for collaborative learning environments. However, especially for individual learning, adaptivity plays a significant role (e.g., for intelligent tutoring systems). Moreover, these systems are not necessarily applicable for longer argumentative writing tasks, which often exist in practical pedagogical environments. Finally, as Dillenbourg (2002b) states in his scripting concepts, scripting approaches might limit the creativity of argumentation in the long term and therefore come with certain limitations to other learning outcomes.

#### 2.3.4.3 Adaptive Support Approaches

Besides the representational guidance approaches and discussion scripting approaches, Scheuer (2015) distinguished between a third class of argumentation learning systems: adaptive support approaches. The idea of adaptive support approaches is mainly based on the concept of intelligent tutoring systems. These approaches aim to use computational approaches to model the argumentation learning process of students and then provide adaptive support, e.g., in the form of text messages that highlight mistakes in a graph. Mostly based on pre-defined rules, these systems are designed to provide pedagogical feedback on a learner's action, gives solutions, hints, and recommendations to encourage and guide future activities in the writing processes, or automated evaluation to indicate whether an argument is syntactically and semantically correct (e.g., Scheuer et al. (2010); Scheuer (2015); Pinkwart et al. (2009); Scheuer et al. (2012a)).

Past approaches are mainly based on syntactical analysis, e.g., using ruled-based approaches. For example, the system LARGO analyzes legal arguments using syntactical analysis (Pinkwart et al., 2009). Based on an ontology, Scheuer et al. (2012a) introduced the system LASAD as an adaptive support system based on graphical representation (see Figure 11 and Figure 12). The LASAD system was a milestone in argumentation skill learning tools, since it combined various different design features, such as sentence openers, different representational guidances, and adaptive feedback messages. The system's adaptivity is based on XML-defined concepts and patterns. Nevertheless, as Scheuer (2015) states, "rigorous empirical investigations of adaptive support for collaborative or argumentative learning are ... rare" (Scheuer (2015), p.120). Also, for the LASAD system, no extensive learner-centered evaluation exists that indicates that the different approaches and the design influence students' argumentation skills in the field.

Moreover, other adaptive argumentation feedback systems exist (e.g., Huang et al. (2016); Suthers and Hundhausen (2001)), but they are usually not based on novel advances in AI and lack a user-centric design along with user studies. Most current systems are based

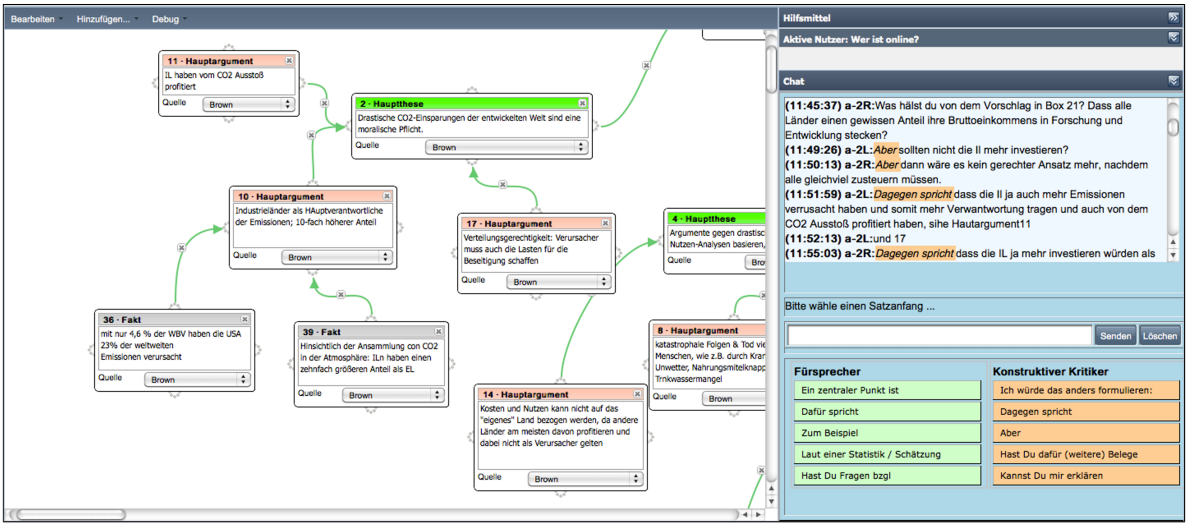



Figure 11: LASAD user interface adapted from Scheuer et al. (2012a): The system combines graph-based representations with sentence openers based on a domain-independent ontology.
Source: Scheuer et al. (2012a)

on expert-defined ontologies, patterns, dictionaries, or syntactical rules and are therefore limited in providing a state-of-the-art personalized learning experience in everyday educational tasks, e.g., when writing a convincing text.

In fact, in other disciplines, such as in NLP, other forms of argumentation system exist. Since 2016, research in NLP has utilized argumentation mining algorithms and has embedded these approaches in simple text input tools. Lippi and Torroni (2016b) developed the first online argumentation mining tool that was made available for a broad audience. Their tool, MARGOT (see Figure 13), is available as a web application and processes a text that is put in the corresponding editor field. Based on the Toulmin Model Toulmin (2003), Lippi and Torroni (2016b) trained an ML model that classifies claim and premises in a given text. After processing and analyzing the text, the results are displayed on the user interface. Claims are displayed in bold font, whereas premises are displayed in italic style (Lippi and Torroni, 2016b).

Also, Lauscher et al. (2019) have provided a simple text-input interface called *Argumin-Sci* that aims to support the holistic analysis of scientific argumentation, including the identification of argumentative components (see Figure 14). Moreover, Chernodub et al. (2019) designed a system to make argumentation identification in text available for individuals. In their system, called TARGER, a user can analyze the argumentative structure of an input text. The results are then presented below the input. Claims are highlighted in red and premises are marked in green (see Figure 15). These approaches originated in the NLP and computer linguistics community and were mostly built with the aim to

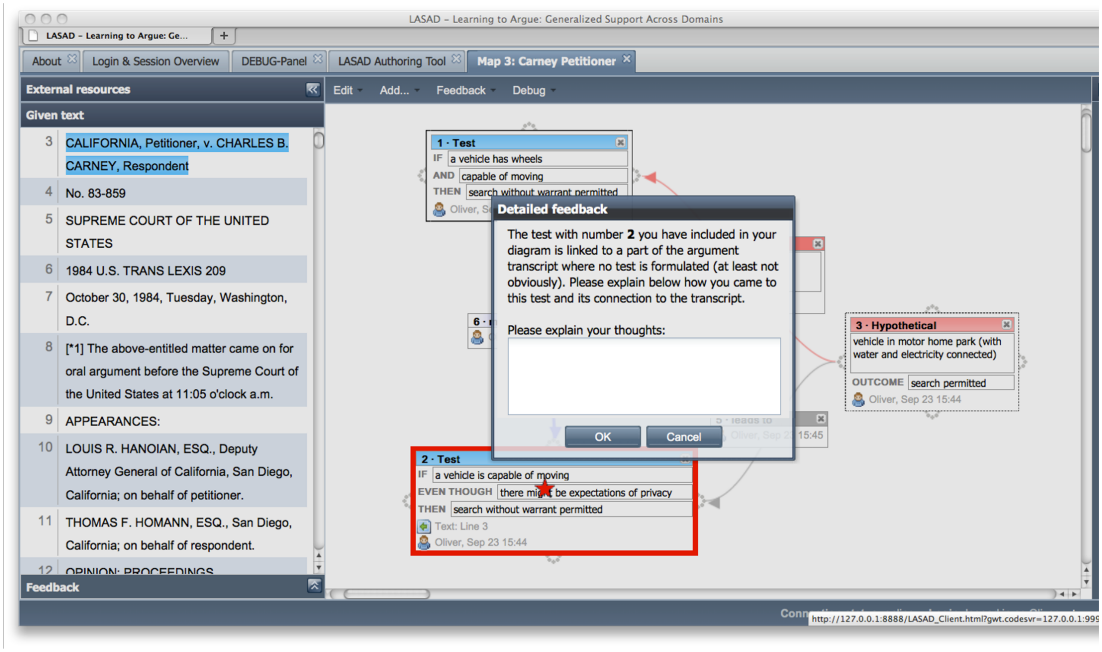



Figure 12: LASAD user interface adapted from Scheuer et al. (2012a): The system provides users with feedback messages based on detected patterns.
Source: Scheuer et al. (2012a)

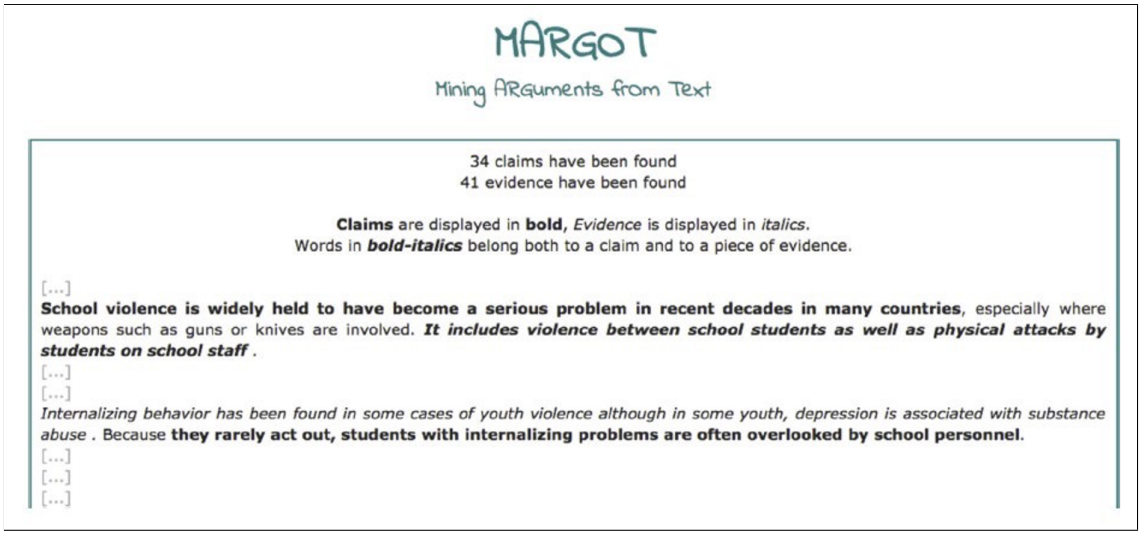



Figure 13: Screenshot of the argumentation mining tool Margot.
Source: Lippi and Torroni (2016b)

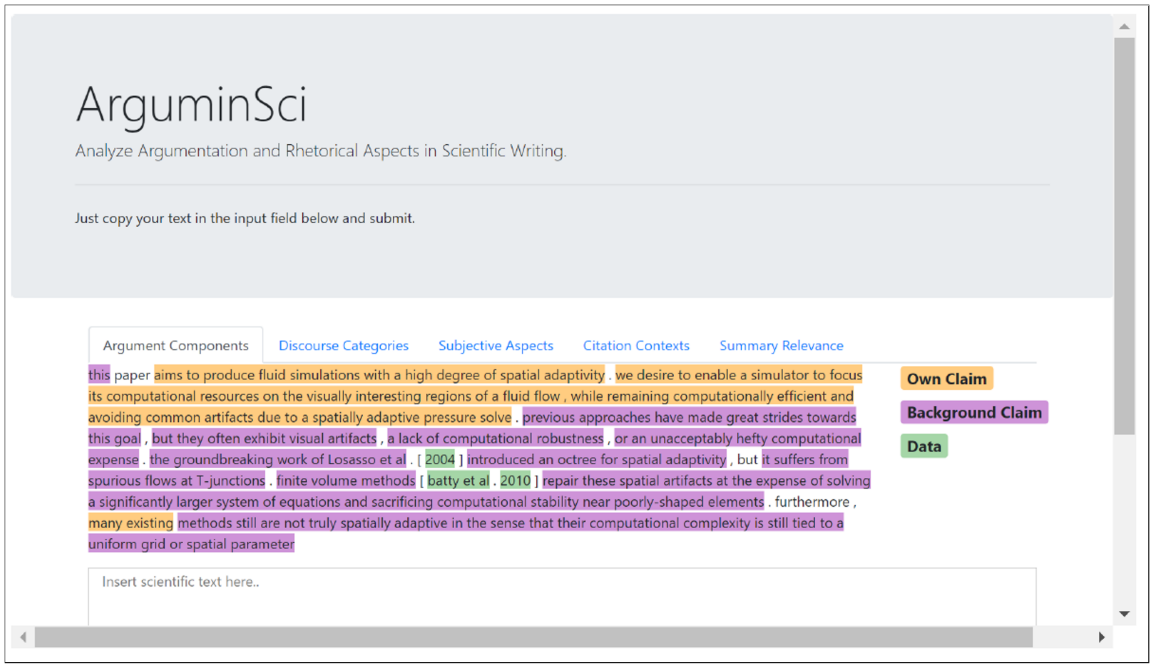



Figure 14: Interface of ArguminSci for argument claim detection in scientific writing.
Source: Lauscher et al. (2019)

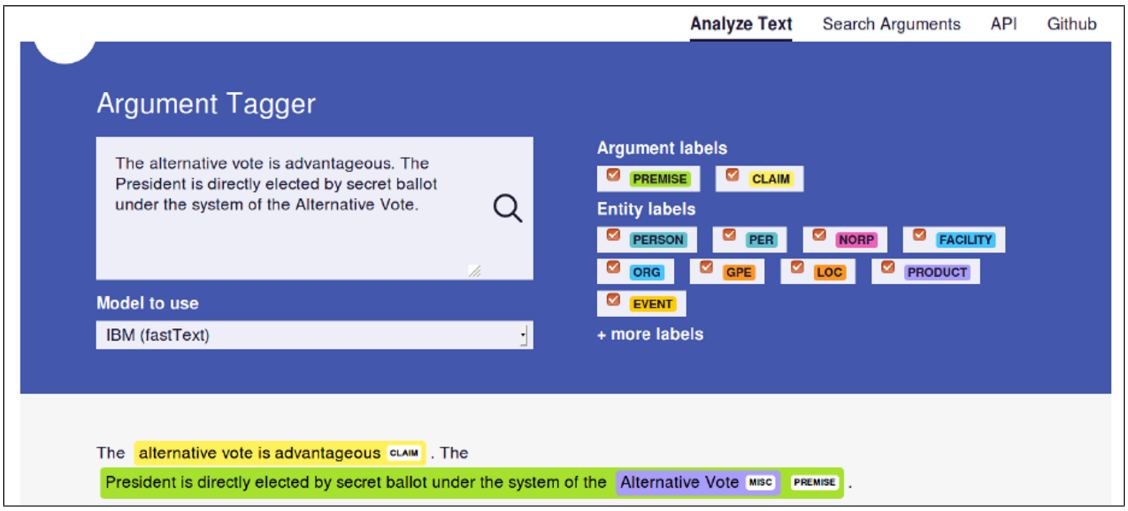



Figure 15: Argument Targer based on different argumentation mining models for claim and premise highlighted.
Source: Chernodub et al. (2019)

make novel argumentation mining algorithms available for non-programmers. Nevertheless, they were not designed for a certain pedagogical scenario, nor with a user-centered perspective. Furthermore, these systems are not taking technology-mediated learning, human—computer interaction, or IS perspectives into account and are thus not necessarily easy to use and easy to access for students in their learning journey. In fact, a leaner would always have to open the website, select a certain model, and then copy his or her text into the input field.

#### 2.3.4.4 AI-based Argumentation Writing Support Systems

With this dissertation, I propose and contribute to two novel classes of argumentation learning systems. These systems are centered around a technology-mediated learning perspective (Gupta and Bostrom, 2009) and aim to tackle the challenges from prior systems by using a more holistic design across research disciplines (human—computer interaction, IS, and NLP).

Both novel argumentation learning system classes not only aim to address the educational task (writing coherent arguments) but are also intended to build a holistic argumentation learning system embedded in a pedagogical scenario (structure) focusing on the learner through a user-centered and theory-motivated design. Moreover, the novel classes of argumentation learning systems are both built using recent advances in AI, which enable an adaptive and individual learning experience irrespective of the instructor, time, and place. Overall, three core properties of the class of AI-based argumentation learning systems are evident, at least one of which is distinct from existing classes (i.e., Table 4):

- **Accessibility:** Built on cloud technology and responsive design, these systems offer wide-ranging accessibility for different user groups as web-based interfaces with transparent explanations.
- **Intelligence:** Leveraging recent advances in NLP and ML, the heart of the novel class is based on AI. This carries the advantages of self-learning systems, which improve and scale to wider use cases based on growing user (data) bases. However, this might eventually produce unintended effects, including a lack of explainability or ethical considerations, which should be addressed with the next property.
- **Socio-technical and user-centered design:** Based on a socio-technical system perspective and an IS view, the novel system class incorporates not only user requirements but also a holistic systems perspective, for example, by considering the

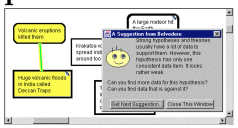
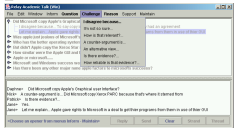
learning task, the technological embedding (e.g., properties of the technology), and the pedagogical scenario (e.g., ethical considerations).

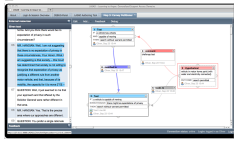
AI-based argumentation writing support systems focus on providing the user with an adaptive argumentation writing interface on which the user receives a self-monitoring and self-evaluation dashboard based on his or her text. More about the specific design principles, the architecture, and the implementation for AI-based argumentation writing support systems can be found in section 5 for persuasive peer reviews or in section 7 for argumentative business pitches. At the time of writing this thesis, no AI-based writing support system for argumentation skills exists that fulfilled all of these three characteristics.

#### 2.3.4.5 Conversational Argumentation Learning Systems

Besides the new system class of AI-based argumentation writing support systems, with this work, I introduce the novel class of pedagogical conversational agents for argumentation skill learning. The class of pedagogical conversational agents adds a completely new perspective and paradigm to past argumentation skill learning research: dialogue-based argumentation learning. As explained in subsection 2.3, conversational learning settings have added a new component to various other skill learning domains (e.g., Cai et al. (2019); Ruan et al. (2019); Winkler et al. (2021)). Conversations and debates represent one of the most natural ways to interact in reasoning and argumentation (Kuhn, 1993; Jonassen and Kim, 2010). Technology-mediated learning might provide a completely new perspective on supporting and facilitating pedagogical scenarios in large-scale learning environments through a dialogue-based learning tool. An adaptive conversational tutoring system provides students with adaptive and instant feedback, theoretical input, and step-by-step guidance during their writing process. Similar to the class of AI-based argumentation writing support systems, the novel class of pedagogical conversational agents come with the characteristics of accessibility, intelligence, and socio-technical and user-centered design. However, it also adds a conversational learning paradigm to the pedagogical scenario. Moreover, pedagogical conversational agents often embed anthropomorphizing design features (e.g., an avatar, a name, or human-like communication characteristics, i.e., Winkler and Söllner (2018); Noetel et al. (2021)). These characteristics have recently been shown to add additional benefits to the interaction with conversational agents in general (Zierau et al., 2020; Feine et al., 2019). At the time of writing this thesis, no pedagogical conversational agent for argumentation skill learning was found in the academic literature that fulfilled all these properties.

In this dissertation, I design, implement, and evaluate different pedagogical conversational agents for argumentation skill learning. Especially section 6 provides deep insights into my findings on how the system class of pedagogical conversational agents for argumentation skill learning helps students to learn independently of an instructor, time, and location.

| Approach | Description | Theoretical Lens | Example References | Limitations |
|---|---|---|---|---|
| **Representational guidance approaches**  | Try to support argumentation learning by providing representations of argumentation structures with the objective of stimulating and improving individual reasoning, collaboration, and learning. One common approach is to represent argumentation discourse in the form of linked graphs. | Representational guidance theory (e.g., Suthers (2003)) | Nussbaum et al. (2007); Pinkwart et al. (2006); Suthers and Hundhausen (2001); Bell (1997) | Focus on the representation of argumentation, not on the learning itself; larger argumentation graphs tend to confuse users and are not useful for longer writing tasks; representations are usually built on domain-dependent ontologies and are thus limited for a wide-scale adoption (Scheuer et al., 2010). |
| **Discussion scripting approaches**  | Aim to provide structured elements for argumentation learning processes to foster interactions based on script theory of guidance. One common approach is to let learners choose from predefined sentence openers when composing new text content. | Script theory of guidance (e.g., Dillenbourg (2002b); Fischer et al. (2013); Kollar et al. (2006)) | Fischer et al. (2013); Jermann et al. (1999); Mcmanus and Aiken (1996); Dillenbourg (2002a) | Missing adaptivity in scripted learning (i.e., Scheuer (2015)); not necessarily applicable for longer argumentative writing tasks; might limit the creativity of argumentation (i.e., through over-scripting Dillenbourg (2002b)). |

| Approach | Description | Theoretical Lens | Example References | Limitations |
|---|---|---|---|---|
| **Adaptive support approaches**  | Aim to provide pedagogical feedback on a learner's action and solutions, hints, and recommendations to encourage and guide future activities in the writing processes, or automated evaluation to indicate whether an argument is syntactically and semantically correct. | Mostly missing (i.e., Scheuer (2015)) | Non-AI-based adaptive systems exist, e.g., Huang et al. (2016); Pinkwart et al. (2009); Scheuer et al. (2012b,a); AI-Based learning systems motivated but not implemented (i.e., Lawrence and Reed (2019); Stab and Gurevych (2017)); argumentation mining Systems exist but without any educational or user-centered embedding, e.g., Chernodub et al. (2019); Lippi and Torroni (2016b); Lauscher et al. (2019). | Current system based on dictionary approaches or syntactical rules; limited in providing adaptive and individual learning feedback; empirical evaluation missing (i.e., Scheuer (2015)); no system based on novel ML and NLP approaches exist; no user-centered adaptive front end. |

| Approach | Description | Theoretical Lens | Example References | Limitations |
|---|---|---|---|---|
| **AI-based Argumentation Writing Support Systems** | Socio-technical system and user-centered design; built on self-learning algorithm based on NLP, ML, and argumentation mining; accessible via the web with transparent insights through explanations and learning recommendations. | Combinations of several proven learning elements/theories, e.g., self-regulated learning, learning control mechanism, or nudging in a holistic systems design. | This dissertation, e.g., section 5 or section 7 | Domain-dependent (depending on the used corpus); large implementation effort necessary before the first application; focus on socio-technical design aspects important. |
| **Conversational Argumentation Learning System** | Same as AI-based argumentation writing support system; natural dialogue-based learning experience with transparent insights through explanations, learning recommendations, representational guidance, and automatically generated feedback based on a student's argumentation skill level. | Same as AI-based argumentation writing support system | This dissertation, e.g., section 6 | Same as AI-based argumentation writing support system |

Table 4: Overview of different IT-based argumentation learning approaches.
Source: Own illustration

## 2.4 Learning Process: Learning to Argue vs Arguing to Learn

To investigate how new types of technology-mediated learning impact argumentation skill learning, the technology-mediated learning framework of Gupta and Bostrom (2009) suggests to look precisely at how students use the IT artifact as part of their learning process. Students' learning processes are regarded as a structuring process in which individuals learn and adapt the structures of the learning method based on their learning objectives (Gupta and Bostrom, 2009). The concept of scaffolding can be applied to describe the process of technology appropriation and focuses on how-to guide or facilitate students in constructing new knowledge (Azevedo and Hadwin, 2005; Oh and Jonassen, 2007; Janson et al., 2019).

In the previous section, I indicated already that the learning process of argumentation can play a threefold role in certain pedagogical scenarios. First, a pedagogical scenario can aim to particularly train argumentation skills by asking learners to participate in argumentative writing activities with the pedagogical focus to promote argumentation skill learning. This is commonly referred to as *learning to argue* (Jonassen and Kim, 2010). Scholars have shown that participation in argumentation activities, such as group discussion on what defines bad or good argumentation, has positive effects on argumentation skills (i.e., Jonassen and Kim (2010); Kuhn (1993, 2005); Jonassen and Land (2012)).

Secondly, practicing active debates and argumentation can be used as a didactical method to achieve other learning outcomes, e.g., to manifest certain factual knowledge. To apply argumentation as a didactical method in a learning process is defined as *arguing to learn* (Jonassen and Kim, 2010; Asterhan and Schwarz, 2016). For example, Asterhan and Schwarz (2009) showed that engaging students in argumentative discussions led to higher learning outcomes in biology. The effective use of argumentation as a means to improve student knowledge and understanding has been widely investigated (i.e., Asterhan and Schwarz (2016); Jonassen and Kim (2010); Scheuer (2015)).

Third, actively engaging in argumentation has been shown to positively impact other metacognition skills such as communication and problem-solving in general (Duschl and Osborne, 2002; Kuhn, 2005; Van Eemeren and Grootendorst, 2016). Logical, structured arguments are a required precondition for persuasive conversations, general decision-making, and drawing acknowledged conclusions. As Kuhn (1992) states, the skill to argue is of great significance, not only for professional purposes like communication, collaboration, and for solving difficult problems but also for most of our daily life. Hence, researchers have used argumentation learning approaches, e.g., for promoting productive civil discourse (Diana, 2018).

In this dissertation, I mainly focus on the learning process of argumentation skill learning (*learning to argue*) and how new classes of technology-mediated argumentation learning artifacts influence the argumentation learning process of students. Specifically, I explore how AI-based argumentation writing support systems help students to improve their metacognitive argumentation skills in a peer review setting in a large-scale lecture (see section 5). Moreover, I investigate how students can be nudged with AI-based argumentation writing support systems to write more convincing entrepreneurial elevator pitches with argumentation self-evaluation (see section 7). Lastly, I also research the design and the effects of conversational argumentation learning systems and how they help students to improve their argumentative writing in a peer-learning exercise (see section 6). As stated, in this dissertation I focus on metacognition skill learning. However, the effects of argumentation skill feedback on other pedagogical outcomes is a strong candidate for future research.

## 2.5 Technical Background

In this dissertation, I aim to investigate how to design novel argumentation learning systems based on recent advantages of AI to support argumentation skill learning independent of an instructor, time and, location, and to answer how these design findings influence students' argumentation skills and learner-experience. Hence, in the course of my research I do not only collect and annotate new corpora for argumentation skill learning based on student-written text, but also design, train, and tune several algorithms based on recent advantages in NLP and ML. Since my work contributes to an interdisciplinary field with publications in IS, human—computer interaction, and NLP, I want to provide a brief technical background on the methods of NLP, ML, corpus collection, and the respective evaluation metrics applied in this work.

### 2.5.1 Natural Language Processing

With the rise of textual data especially through the Internet and the available computing power, the potential of NLP for automated knowledge extraction out of unstructured or semi-structured data gained growing importance. The goal of NLP is to mathematically analyze human language, understand and interpret the context to derive information and knowledge, and enable human-machine-interaction (Johnson et al., 2015). In this work, NLP algorithms are used to assess and analyze the skill level of students, e.g., the argumentation skill level through argumentation mining.

Following Hashimi et al. (2015), the application of NLP for ML can be divided into

three major steps: preprocessing, feature generation, and feature selection. As Yang and Pedersen (1997) stated, feature selection actually compromises the step of dimensionality reduction of features and thus, can be called feature reduction.

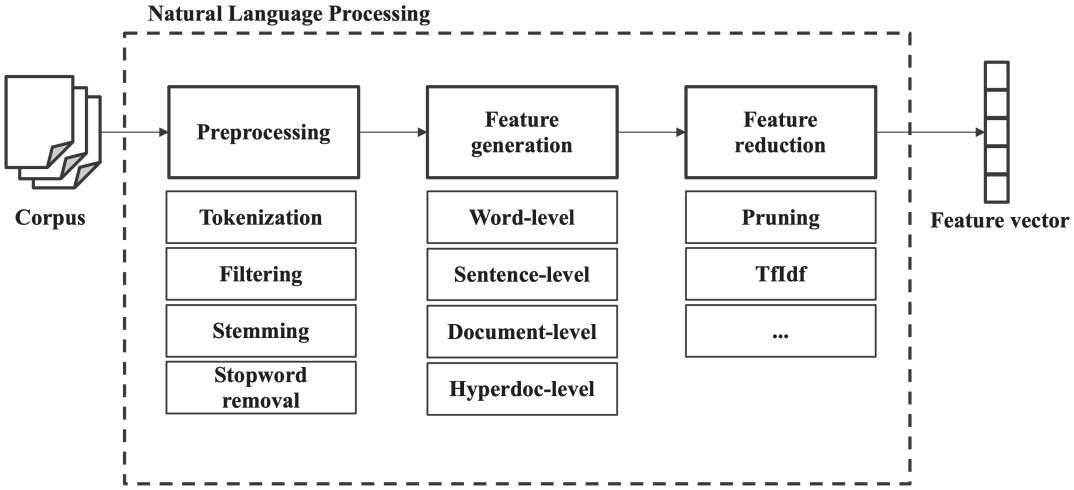



Figure 16: Overview of NLP techniques as standard text processing for ML-based systems. Source: Own illustration

All in all, the objective is to prepare data for the modeling in order to apply ML algorithms. In the following, the standard preprocessing steps for ML will be briefly described according to the three steps: preprocessing, feature generation, and feature reduction (see Figure 16). More information on these general steps can be found in Bird et al. (2009).

The NLP flow starts with different preprocessing techniques, which take a plain text object and transform it into an array of tokens. Specifically, in NLP, but also in general Data Science, preprocessing steps are of special significance for the overall outcome of a ML system (Uysal and Gunal, 2014). The standard preprocessing algorithms are mainly used to filter noise such as the elimination of common words or unnecessary text structures (Hammouda and Kamel, 2004).

Using the right features and eliminating noise is extremely important for the performance of most learning algorithms, also related to the widely known curse of dimensionality. More about the curse of dimensionality can be found in (Pestov, 2013). Andrews and Fox (2007) distinguished between four preprocessing steps: tokenization, filtering, stemming, and stopword removal. The steps are briefly explained in the Table 6.

According to Johnson et al. (2015), NLP algorithms rely on processing data along four major dimensions of language: morphology, lexicography, syntax, and semantics. Basically, the dimensions describe the deepness of NLP understanding from the occurrence of single

| Name | Explanation |
|---|---|
| **Tokenization** | Process of splitting a text into individual tokens (Moschitti and Basili, 2004). |
| **Filtering** | Filtering means to remove noise or elements which are considered as not relevant for the text. Those elements could be punctuation, digits, or any other form of special characters (Hammouda and Kamel, 2004). |
| **Stemming** | Reducing words to their base form (stem) by cutting of the suffixes, e.g. the word stem of "answered", "answers" and "answering" is *answer* (Porter, 1980). |
| **Stopword removal** | Process of removing all terms which are thought not to convey any additional information but often occur in a text, e.g. remove "the", "in" and "to". |

Table 5: NLP preprocessing steps for machine learning-based IS.
Source: Andrews and Fox (2007)

characters, the structure of words, the formation of sentences, and finally the meaning of expressions. These dimensions are illustrated in Figure 17 with simple examples.

Morphology deals with the formation of words. It studies how words are put together, e.g. which syllables, pre-or suffixes are used in a word. In general, words with more syllables can be considered as more sophisticated. According to Gunning, their usage displays a more elaborated expression of language (Gunning, 1969).

The lexical level focuses on the structure of words and word classes. Whereas, syntax deals with the formation and grammar of sentences. It examines how words form a sentence and, which role they take in the sentence, e.g. determining the subject, object, or verb in a sentence (Andrews and Fox, 2007).

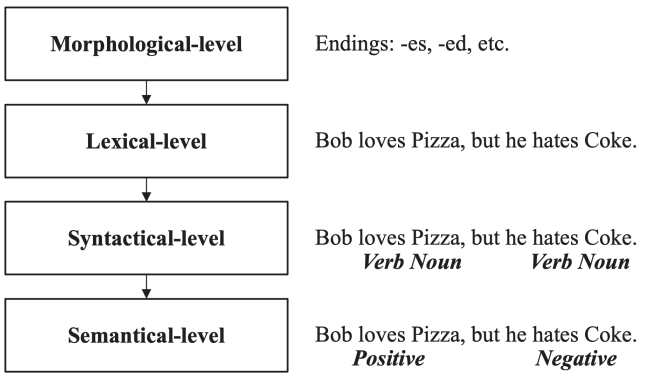



Figure 17: Overview of the levels of linguistics illustrated with examples.
Source: Own illustration

Finally, the semantic level goes beyond the structure of words and sentences. It deals with

the meaning of words or sentences, e.g. determining the sentiment of words. This can be considered the first step of natural language understanding. However, this level is thought to be the most challenging in linguistic analysis (Shahaf and Amir, 2007). An example of semantic analysis is the determination of the sentiment of posts and classifying them in positive or negative (Pang and Lee, 2008). More insights into NLP feature engineering for ML-based systems and a classification of NLP-text features will be investigated in subsection 4.1.

### 2.5.2 Machine Learning

ML is described as "the area of computer science focusing on the development and implementation of systems that improve as they encounter more data" (Pustejovsky and Stubbs, 2013). Michie et al. (1994) define ML as an "automatic computing procedure based on logical or binary operations, that learns a task from a series of examples".

The objective of ML is to generate classifying expressions simple enough to be easily understood by humans. Algorithms must learn and adopt human reasoning to provide insights into the decision process. Like statistical approaches, comprehensive background knowledge may be exploited in development, however, in contrary to statistical methods, ML usually does not require manual adjustment and structuring of the given problem (Michie et al., 1994).

When working with NLP, ML can be mostly divided into three major categories:

- **Supervised learning:** Usually referred to as classification, describes a function mapping the input data to a fixed set of priorly human-defined labels.
- **Unsupervised learning:** Also called clustering, comprises a method that tries to identify unknown structures from an input set of unlabeled data.
- **Semi-supervised learning:** Combination of clustering and supervised learning, where input data is mapped to labeled and unlabeled data (Pustejovsky and Stubbs, 2013).

Moreover, Mohri et al. (2013) state that ML algorithms can be categorized according to the character of their learning into four additional categories: transductive learning, online learning, reinforcement learning and active learning. However, since this work mostly applies supervised ML algorithms, I will only provide a brief introduction the the categories of supervised learning algorithms.

### 2.5.3 Supervised Classification Approach

Supervised learning can be distinguished into labeling data with discrete or continuous values. Searching continuous target values is called a regression problem, e.g. determining the income class of potential customers. If discrete values are used, it is referred to a classification problem, e.g. identifying if a text passage consists of a positive or negative message (Pang and Lee, 2008). However, all regression problems can be transformed to classification problems by defining classes for the numerical values, e.g. income classes from *"0 - 1000€, 1000 - 3000€ and 3000€ +"* Bird et al. (2009).

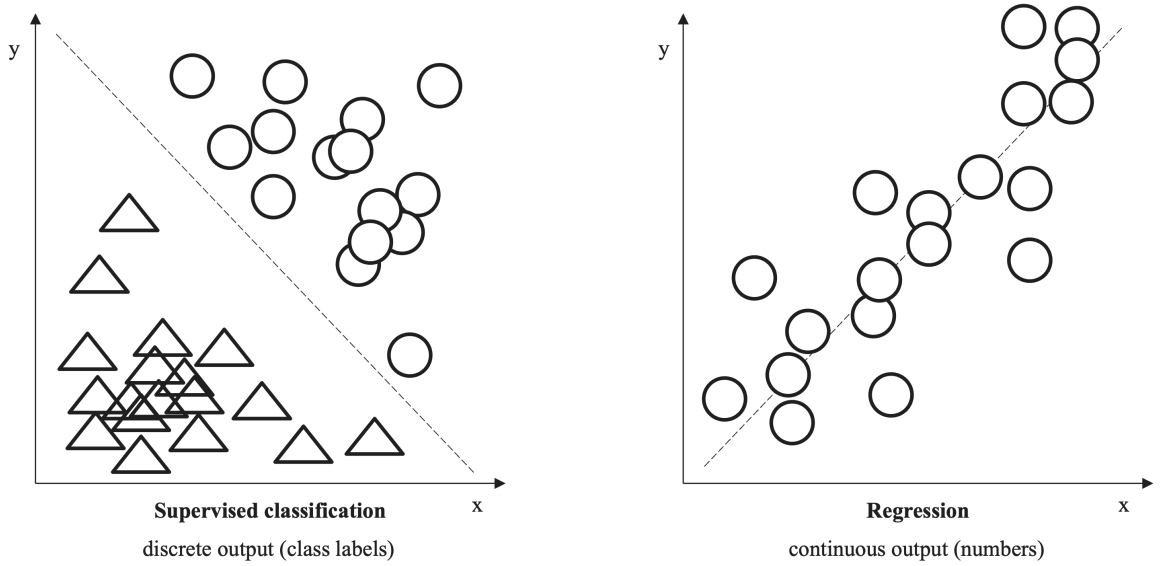



Figure 18: Structure of available data for supervised classification.
Source: Bird et al. (2009)

Hence, choosing between classification and regression depends on the output type of the learning model. If the objective is to generate class categories and label the input data clearly into class A or class B, a classification algorithm is applied. However, if the output is rather continuous between two categories, also a regression with a scoring could be applied, e.g. the input is 80 percent class A and 20 percent class B. All in all, the right selection between regression and classification also depends on the applied domain and the overall model process, e.g. if human monitoring is involved in the process or not.

In general, the framework of a supervised classification is divided into a training and a prediction phase displayed in Figure 19.

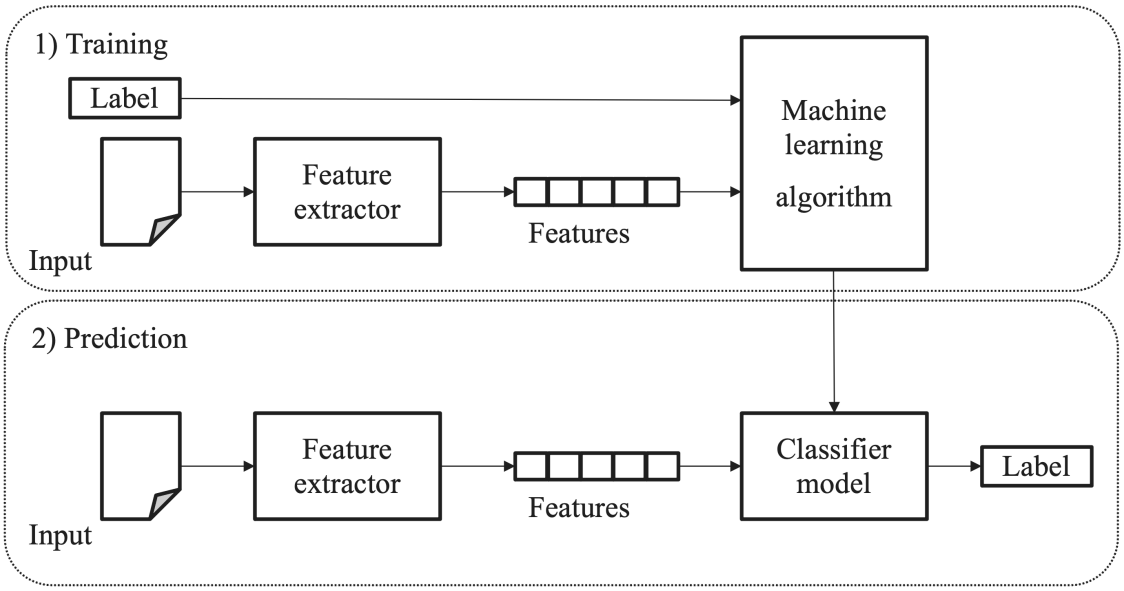



Figure 19: Overview of supervised classification framework with training and prediction. Source: Bird et al. (2009)

During training, the feature extractor preprocesses the manual labeled input data. The corpus is transformed to numerical values comprising the steps of preprocessing, feature generation, and feature reduction. In fact, the actual feature generation process is probably the one of most significant for the performance of the classifier. The right choice, design, and implementation of feature representations algorithms along with word-, sentence-, document-, and hyper-document-level features are of special interest for the optimization of a classifier model (Uysal and Gunal, 2014).

The processed feature sets capture the most relevant information about the input data. Afterward, pairs of feature sets and labels are fed into the ML algorithm to generate a model. The model is then trained and evaluated with all available labeled data. When training is finished, the actual prediction phase starts. Here, the same feature extractor is used to convert unseen inputs to feature sets. The already trained model then uses the feature sets to predict new unseen data.

Nevertheless, there are general limits on how many features one should use during the training phase. If too many features are used, the training model will fit too well to the training data and, if new data is added in the prediction phase, the model does not perform well on the unseen data. This is called overfitting. On the other side, if too few features are provided in training, the classifier does not have enough examples to distinguish the labels.

Therefore, Bird et al. (2009) suggests applying an error analysis phase with a development

| Categories | Example of Learning Algorithms |
|---|---|
| **Decision Tree Learning** | Decision Tree Classifier, Random Forest Classifier |
| **Bayesian Learning** | Naive Bayes, Bernoulli Naive Bayes, Gaussian Naive Bayes |
| **Instance-based Learning** | K-Nearest-Neighbor |
| **Deep Learning** | Single Perceptron, Multilayer Perceptron |
| **Linear Classifiers** | (Linear) Support Vector Machine, Logistic Regression |
| **Ensemble** | Bagging, Boosting |

Table 6: Non-exhaustive overview on classifier learning categories.
Source: Michie et al. (1994); Opitz and Maclin (1999)

set. The available corpus is split into a *development* and a *test*, also called evaluation set. The development set is then split again into a *training* and *dev-test set*. The *training* and *dev-test set* are used for training, parameter optimization of the classifier, and error analysis. Finally, the unseen test data is then used to simulate a prediction of new data to avoid the risk of overfitting.

Michie et al. (1994) named four basic supervised ML types, which form the standard categories for classification algorithms: Decision Tree Learning, Bayesian Learning, Instance-based Learning and Artificial Neural Networks (Michie et al., 1994). Since 1994 other algorithms emerged such as Support Vector Machines (linear classifiers) (Cortes and Vapnik, 1995) and multi-combined algorithms, which are basically estimators that combine several classifiers to an ensemble (Opitz and Maclin, 1999). An non-exhaustive overview of the categories can be found in Table 6.

It must be mentioned, that comparing these algorithms strongly depends on the domain, the used data and therefore, a generalization is not easy and full of pitfalls (Salzberg, 1997). More on ML and NLP can be found in (Michie et al., 1994; Bird et al., 2009).

### 2.5.4 Corpus Design

The performance of classification algorithms can only be as valid as the data they represent. Hence, it is significant to devote special attention not only to the preprocessing techniques of NLP algorithms but also to the raw data they are fed with. Especially, when a new corpus is collected it becomes necessary to devote attention to the principles of corpus design. Therefore, this section will briefly summarize the core concepts for designing a corpus.

**Definition Corpus:** *A corpus is a collection of machine-readable texts that are representatives of natural human language (Pustejovsky and Stubbs, 2013).*

According to Pustejovsky and Stubbs (2013), there are two major characteristics, which must be considered when constructing a corpus: representativeness and balance with respect to the genre and language that the corpus seeks to represent.

**Definition Representativeness:** *A corpus can be considered as representative of a domain if it represents the full range of variability of that domain (Biber, 1993).*

**Definition Balance:** *The proportions of different kinds of text, a balanced corpus contains, should correspond with informed and intuitive judgments (Sinclair, 2005), e.g. a corpus aiming to represent the "German language" would have to encompass all types of written and spoken texts, from newspaper articles to web forums to books, or television subscripts.*

Hence, researcher usually developed precise guidelines for the creation and especially the annotation linguistic corpora (e.g., Sinclair (2005); Stab and Gurevych (2014a, 2017); Carlile et al. (2018)). In this vein, the objectiveness, also called gold standard, of the corpus is defined and usually also tested based on an annotation study (Stab and Gurevych, 2017).

Certainly, the concepts of representativeness and balance are not easy to define and whether a corpus can ever be truly fulfilling these criteria is not agreed on (Pustejovsky and Stubbs, 2013). The size of a corpus is also important. In general, a corpus can never be large enough in order to fulfill the two design principles. NLP and ML algorithms are based on statistical methods. In statistics, the law of large numbers applies, which guarantees stable long-term results for the average of some random events. The more data a corpus consists of, the more it achieves the goal of representativeness and balance (Pustejovsky and Stubbs, 2013).

However, in practice, the corpus size is limited by time, money, resources, and attention span (Pustejovsky and Stubbs, 2013). In fact, the most important characteristic a corpus should hold is described as: "all expected and relevant phenomena that are expected should be represented" (Pustejovsky and Stubbs, 2013). If a single test set consists of several closely related features, such as features drawn from a single document, the size of the test set should be increased to ensure that this lack of diversity does not skew the evaluation results (Pustejovsky and Stubbs, 2013).

After the corpus is collected, the documents are processed with NLP and fed as feature vectors to ML classifiers.

### 2.5.5 Evaluation Metrics

In this section, an overview of the ML algorithm elevation metrics used in this work will be provided. The different terms will be shortly defined. The basic representation for the evaluation is the so-called confusion matrix:

**Definition Confusion matrix:** *A confusion matrix displays each cell [i,j] indicating how often label j was predicted when the correct label was i. Hence, the diagonal values indicate labels that were correctly predicted, and the off-diagonal values indicate false predictions (Bird et al., 2009).*

The performance of ML algorithms is measured and compared with three different metrics: accuracy, precision, and recall firstly used by Kent et al. (1955). Moreover, the so-called Fß-score is calculated to provide a weighted average for precision and recall.

**Definition Accuracy:** *Accuracy describes the systematic error of a classification. It is defined by:*

$$accuracy = \frac{truePositives + trueNegatives}{(truePositives + trueNegatives + falsePositives + falseNegatives)}$$

(van Rijsbergen, 1979)

**Definition Precision:** *Precision is the proportion of retrieved documents which are relevant. It is defined by:*

$$precision = \frac{truePositives}{(truePositives + falsePositives)}$$

(van Rijsbergen, 1979)

**Definition Recall:** *Recall is the proportion of relevant documents retrieved. It is defined by:*

$$recall = \frac{truePositives}{(truePositives + falseNegatives)}$$

(van Rijsbergen, 1979)

**Definition Fß-score:** *Fß-score provides a weighted mean for precision and recall. The ß of the Fß-score can be adjusted in order to increase or lower the influence of either precision or recall. For ß=1 an equally balanced mean between precision and recall is calculated. By setting the value ß=0.5, the emphasis of precision is increased and vice*

*versa, by setting ß=2 the emphasis of recall is increased. It is defined by:*

$$F\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

(Baeza-Yates and Ribeiro-Neto, 1999)

### 2.5.6 Argumentation Mining

This dissertation builds on recent advances in NLP and ML, particularly argumentation mining. Argumentation mining is a new research field in computational linguistics that focuses on the extraction and analysis of arguments from textual corpora as well as on following and analyzing the lines of argumentation (i.e., the interplay between arguments).

Since 2007, scientists have published studies on argumentation mining in legal texts, online reviews, or debates (Mochales and Moens, 2011; Palau and Moens, 2009). Argumentation mining analyzes the arguments of a given text based on a defined argumentation structure (often based on the Toulmin model explained earlier (Toulmin, 1984)). Argumentation structures can be identified on four different levels (i.e., Lawrence and Reed (2019)):

1. **Argument identification**: identifying the argumentative parts in raw text and setting up its boundaries versus a non-argumentative text (e.g., Florou et al. (2013)).
2. **Argument component classification:** the primary purpose of this subtask is to classify the components of the argument structure. Classifying an argumentative text into claims or premises is one popular way of tackling the target of this subtask (Mochales and Moens, 2008; Stab and Gurevych, 2014a; Lawrence and Reed, 2019).
3. **Argumentative discourse analysis:** this task's purpose is to identify the discourse relations between the various components existing in the argument. A typical example of this subtask is the identification of whether a support or an attack relationship exists between the claim and the premise (Palau and Moens, 2009).
4. **Argumentation quality assessment:** accessing the quality of a certain argument component (e.g., is a claim sufficiently supported by the premises or how relevant are the premises for the particular claim) (Carlile et al., 2018). These qualitative attributes are usually annotated by a human on a Likert scale. In the past, they included relevance, strength, or clarity of claims and premises (e.g., Carlile et al. (2018); Lawrence and Reed (2019)).

Also, argumentation mining researchers have become increasingly interested in intelligent

writing assistance (Song et al., 2014; Stab and Gurevych, 2017), since it enables adaptive argumentative writing support with tailored feedback concerning arguments in texts. However, the complexity of this technology's use in pedagogical scenarios for educational purposes has hitherto been poorly assessed (Lawrence et al., 2004).

Despite the availability of literature on argumentation systems, broader insights into system design and development, particularly for the new class of AI-based argumentation learning systems, is scarce (Afrin et al., 2021; Scheuer, 2015; Wang et al., 2020). Although existing studies have extended the understanding of technology-mediated argumentation learning, they were usually conducted to answer specific research questions rather than to obtain information that could be used for widespread application (e.g., Afrin et al. (2021); Chernodub et al. (2019); Wang et al. (2020); Lippi and Torroni (2015b); Lawrence and Reed (2019)). As Stab and Gurevych (2017, p. 649) state, "It is still unknown, however, if feedback provides an adequate guidance for improving students' argumentation skills. To answer this question, it is required to integrate the proposed model in writing environments and investigate the effect of different feedback types on the argumentation skills of students in future research". Scheuer (2015, p. 126) also notes, "Rigorous empirical research with respect to adaptation strategies is almost absent; a broad and solid theoretical underpinning or theory of adaptation for collaborative and argumentative learning is still lacking".

Although research has explored the development and use of argumentation learning systems, it has often neglected to present the general design principles and implementation, particularly for the novel class of AI-based argumentation learning systems, for further developments in the field. In this dissertation, I aim to fill this gap using three research activities: (1) formulation of design principles, (2) design and development, and (3) demonstration and evaluation of the design findings in different contexts and domains. In order to do so, I use a set of qualitative and quantitative research methods, which I will explain in the next section.

# 3 Methodological Background

*"The true method of knowledge is experiment."*

- William Blake

As the quote of English poet William Blake indicates, the process of truth-finding and knowledge generation can be based on different underlying schools of thought and methodologies. Hence, in this section, I not only aim to outline the necessary methodological background used in this dissertation, but I also discuss the underlying research epistemology that influenced my research. In particular, I provide the background on conducting a systematic literature review and building a taxonomy (these methods are mainly applied in section 4). Moreover, I will provide an overview of the used qualitative and quantitative methods, such as for user studies, laboratory and field experiments (these methods have been mostly used in section 5, section 6, and section 7. Next, I will introduce the concept of design science research (DSR), which served as an underlying methodological support for the overall organization of my research. Finally, I will briefly introduce the CRISP-DM methodology, which guides the parts of my dissertation to data science.

## 3.1 Underlying Research Epistemology

Research epistemology deals with the question of how to establish scientific theories and distinguish them from wrongly stated propositions. Hence, this short section aims to provide an understanding of how knowledge and theories are formed and validated in this work. Therefore, I will explain how the different perspectives of epistemology have influenced the applied research methodology of my work.

According to Recker (2013), research in IS is usually influenced by two distinction sciences: natural and social sciences. While natural sciences deal with natural phenomena, social sciences deal with societies and the relationships among individuals in these societies. Recker (2013) states that natural sciences are often considered the "more exact" sciences. Whereas in natural science, for example, a theory would be about the relativity of time and space, a theoretical explanation in social science would be more about how individuals use and adopt technology in a specific context. Also, Döring and Bortz (2015) distinguish between what they call the non-empirical and the empirical sciences. Accordingly, non-empirical sciences consist of formal sciences (e.g., mathematics, philosophy, etc.) and humanities (e.g., technology, legal sciences, history, linguistics, etc.). The empirical sciences consist of social sciences (philosophy, medicine, sociology, etc.),

natural sciences (physics, chemistry, biology, etc.), and engineering sciences (mechanical engineering, electric engineering, computer science, etc.).

Research in IS, in technology-mediated learning, and in human—computer interaction can be considered multidisciplinary, i.e., combining methodologies and paradigms from social sciences and from engineering sciences (Recker (2013); Döring and Bortz (2015)). For example, when designing a novel pedagogical scenario for argumentation skill learning based on an AI learning tool, qualitative design methods for the tool's architecture can be used alongside methods from engineering and formal sciences to develop and train ML algorithms. Last but not least, the novel pedagogical scenario can be empirically evaluated based on psychological constructs. In this dissertation, I utilized methods from both the engineering discipline (e.g., NLP and ML algorithms) and the social sciences (e.g., laboratory experiments) to design novel argumentation learning scenarios based on AI and to investigate their effects on students' argumentation skill learning.

The epistemology in sciences deals with the question of how theories can be created on the basis of data. This data is usually gained through studies with the overall goal to contribute with "truth" to the existing body of knowledge. To achieve a certain finding from data, scientists usually differentiate between inductive and deductive reasoning (Döring and Bortz, 2015). For inductive reasoning, the epistemological process would start with specific data from which patterns and insights are defined to explain a certain phenomenon that could form a theory. Hence, an inductive research process typically aims to establish realizable theories based on specific observations from data. Inductive reasoning is, for example, found in qualitative research approaches to build grounding theories (Urquhart, 2017). Deductive reasoning usually starts from a theoretical angle, e.g., to derive and motivate certain hypotheses based on theory, which are then tested with empirical data (e.g., an experiment). This type of research process is usually found in quantitative research approaches. Additionally, an abductive reasoning approach starts with an observation and seeks to draw a simple and the most likely conclusion from the observation. In contrast to inductive reasoning (looking at data patterns), abductive reasoning builds theories based on untraceable values in the data and hence yields a plausible conclusion but does not positively verify it. These values can then build the foundation for the formation of explanatory hypotheses and theoretical advances (Döring and Bortz, 2015).

Last but not least, theorizing through designing and doing is a more distinctive approach to theoretical thinking in the IS and also in the human—computer interaction disciplines (Mueller and Urbach, 2017). Through "real-world" interventions, researchers seek to change the relationships among variables that are the subject of the theory-building process. When an intervention affects the theoretically desired outcome variables, theorizing

through design and action helps to better understand the underlying phenomenon (Jones and Gregor, 2007)

Following the short outline of epistemological perspectives, the research in this dissertation mainly uses interdisciplinary design science research methods to contribute to the literature gaps and to the stated research questions in technology-mediated argumentation learning (see section 1). By following a design science research approach, I apply both quantitative and qualitative methodologies. For instance, qualitative methods are applied to gain a deeper understanding of argumentation learning processes and technology-mediated learning interventions. These findings are then incorporated in the design and implementation of IT artifacts. Afterward, quantitative evaluations are used to evaluate the design artifacts in large-scale learning environments. Based on constructs from social science, the interplay and the usage of students' learning process is captured with surveys.

Hence, my dissertation is not categorized by one specific epistemological paradigm. In fact, the very objective is to base the research on multiple perspectives and methodology to build a coherent and integrated solution to the proposed research challenges for technology-mediated argumentation learning. To provide better guidance to the reader, I will shortly explain the applied research methodologies in the following paragraphs.

## 3.2 Literature Review

According to vom Brocke et al. (2015), literature reviews are a necessary foundation to structure the knowledge at the beginning of every research endeavour. Especially for interdisciplinary research challenges, a literature review can help to structure the existing concepts, findings and can help to identify unsolved literature gaps (vom Brocke et al., 2015). The field of educational technology and technology-mediated learning overlaps with different research disciplines, such as education, psychology, computer science, and IS (Gupta and Bostrom, 2009; Benbasat and Zmud, 2003). Interdisciplinary research comes with the difficulty that relevant findings might be accessible only in certain editions and databases, which makes it difficult to keep track (vom Brocke et al., 2015). Moreover, different communities often have different terminologies and definitions.

The aim of literature reviews is to identify, analyze, and synthesize existing literature systematically, transparently, and reproducibly (vom Brocke et al., 2015). Webster and Watson (2002) states that a literature review is not only about summarizing existing research findings but also about identifying literature gaps which serves as input for developing a research agenda and theoretical propositions. Hence, a literature review aims to provide an overview of the current body of knowledge by searching for relevant material

in knowledge databases. The literature is found through key words and search terms that represent the phenomenon of interest (Webster and Watson, 2002). Especially for interdisciplinary research, the choice of keywords and relevant synonyms has a significant impact on the search results. Therefore, its important to focus on concepts rather than certain authors to ensure that new streams of literature are also incorporated (Webster and Watson, 2002).

In this dissertation, I conduct multiple systematic literature reviews to access the broad area of technology-mediated argumentation learning systems. First, I conduct a systematic literature review following (Webster and Watson, 2002; vom Brocke et al., 2015) to analyze the foundation of NLP-based ML for adaptive IT artifacts. The findings are summarized in a taxonomy according to Nickerson et al. (2013) in subsection 4.1. Next, I analyze the current state of argumentation learning systems and argumentation mining in section 5. My literature reviews usually follow a five-step process (Cooper, 1988; Arksey and O'Malley, 2005). They consist of (1) *identifying the initial research questions*; (2) *identifying relevant studies*; (3) *selecting the studies*; (4) *collecting the data*; and (5) *compiling, summarizing, and reporting the results.*

## 3.3 Taxonomy Development

Researchers employ classification in the form of taxonomies to structure the underlying qualities and dimensions of a research topic in order to better understand and classify it. Taxonomies aid in the classification of objects based on their similarities and differences, allowing researchers to describe, comprehend, and study phenomena of interest (Nickerson et al., 2013). As Kundisch et al. (2021) states, "taxonomies have a long history in the natural and social sciences and play a central role in the IS discipline, especially given the speed of socio-technical progress that requires continuous efforts of understanding." (Kundisch et al. (2021), p. 1)

Fundamentally, object classification is a cognitive assistance that allows humans to manage the complexity created by multiple variables of interes (Hambrick, 1984; Nickerson et al., 2013). Different approaches to developing categories exist, with taxonomies being the most well-known in IS (Bailey, 1994). A taxonomy is a theoretical study of classifications that can include a multitude of subjects including principles, techniques, and regulations (Bailey, 1994; Usman et al., 2017). Although taxonomies were originally formed on the basis of empirical evidence, they are now also based on conceptual evidence (Nickerson et al., 2013). As a result, taxonomies do more than just classify items; they also facilitate information exchange, provide a better understanding of object interrelationships, and hence aid decision-making (Bailey, 1994; Kundisch et al., 2021).

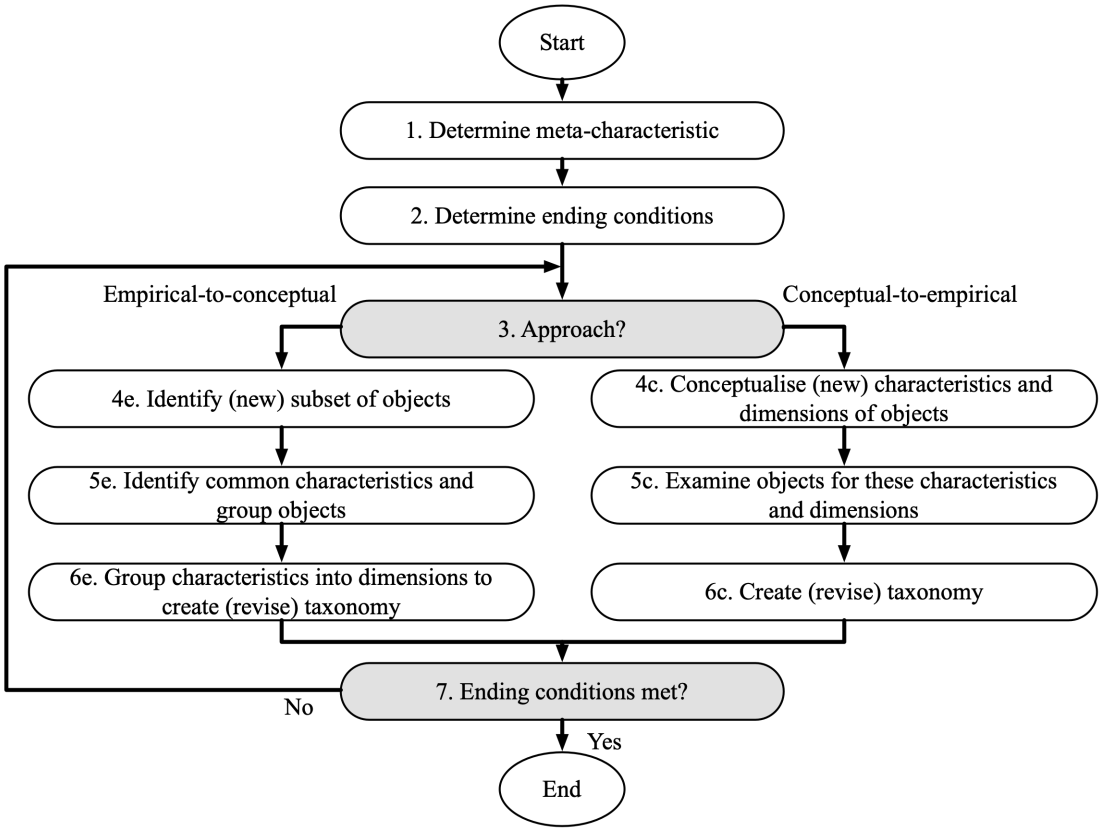



Figure 20: Iterative taxonomy development process.
Source: Nickerson et al. (2013)

Most prominently in IS research, the methodological approach presented by Nickerson et al. (2013) has been applied to structure different phenomena of interest, e.g., for conversational agents (Zierau et al., 2020). For constructing taxonomies based on theoretical foundations (deduction) and empirical evidence (induction), the approach follows an iterative and systematic procedure (see Figure 20 for illustration). By applying the method of Nickerson et al. (2013), research develops different dimensions and qualities based on both published studies obtained in a systematic literature review and empirical evidence of specific meta-attributes.

The definition of a specific phenomenon of interest, also known as a meta-characteristic, is usually the first step in the construction of a taxonomy. All dimensions and qualities should be created with the goal of contributing to this meta-characteristic Nickerson et al. (2013) suggest different subjective and objective criteria, also called ending conditions, which a taxonomy has to fulfil after the iterative taxonomy development process. The following ending conditions (EC) are applied to determine when to terminate the iterative process:

- At least one object is classified under every characteristic of every dimension.

- No new dimension or characteristic has been added in the last iteration.
- Dimensions and characteristics are unique and are not repeated.
- Every known object is classified in the taxonomy.

All ending conditions should be fulfilled by the final taxonomy. I apply the methodological approach of Nickerson et al. (2013) in subsection 4.1 to structure and classify the manifold characteristics and dimensions of NLP features for ML-based information systems. Based on the resulting taxonomy, researchers and practitioners are supported when designing, evaluating, or comparing NLP-based IS, such as argumentation writing support systems or conversational agents. For this reason, the taxonomy also provides the technical foundation and overview of my research endeavour to design and build novel AI-based argumentation learning systems in the subsequent sections.

## 3.4 Quantitative Methods

Quantitative research methods encompass a wide variety of different approaches. My dissertation mostly focuses on experimental studies. Different research methods come with different advantages and disadvantages. Hence, the correct method choice is usually closely related to the posed research question. Mcgrath (1995) suggest three properties to better evaluate and select the right research methodology:

- **Precision**: a high precision in research methodologies is defined as the concept according to which baseline variables can be controlled and measured. Moreover, precision refers to the extent to which variables might be manipulated and to which degree confounding variables can be eliminated.
- **Generalizability**: the maxim of generalizability refers to the concept of how representative the results of the investigated subset of participants is compared to the general population of the real-world phenomena. The more the results are based on generalizable data, the better the research phenomena represent the truth.
- **Existential Realism:** existential realism refers to the concept of how related the participants' observed behavior and their characteristics are to realistic conditions.

It is important to state that there is no method that is able to fulfill all three maxims at the same time. In fact, we can either make optimal use of one of the dimensions and reduce the other two or achieve a balance between two of the dimensions and fail to a

greater degree regarding the third dimension (Mcgrath, 1995). In this dissertation, I use qualitative methods in different contexts, attempting to satisfy the stated properties of Mcgrath (1995). Hence, I apply laboratory experiments (e.g., in section 5 or in section 6) to maximize precision by controlling external variables and confounding conditions. Nevertheless, these laboratory experiments often lack generalizability due to limited sample sizes and participants' realism. Therefore, I conduct field experiments in which I implement my AI-based argumentation writing support system in the field over a period of time to provide realistic conditions with more generalizable observations (e.g., in section 5 or in section 7). In the following subsection, I will introduce the reader to the experimental research methods I used in this work.

In general, experimental studies are applied when the research goal is to provide rigorous and controlled empirical evidence for a certain hypothesis, e.g., to prove a theory. Research that applies experimental research design usually aims to examine cause-and-effect relationships between different variables. Hence, different conditions are applied in a treatment group. To control the effects, a control group is usually used that consists of participants who do not receive a certain treatment. Therefore, the research scholar ensures that the manipulation of the independent variables is controlled for confounding variables that might affect the study (Recker, 2013).

Nevertheless, when designing an experimental study, different choices have to be made. According to Kirk (1982), behavioral experimental studies mostly depend on four criteria. First, the random selection of participants is important to ensure the validity of the results. Kirk (1982) name the measurement of covariates as the second criterion, the implementation of the treatment as the third one, and pre- and post-tests as the fourth one. The most basic form of an experimental study consists of a randomized allocation of participants to a treatment group and one control group. In this vein, one independent variable can be manipulated with an appropriated treatment (Kirk, 1982; Recker, 2013). In technology-mediated learning this could comprise a treatment group of students using an IT artifact providing individual feedback based on AI and one control group using a similar tool for the exact same task but only receiving general feedback (e.g., general rules on what to improve in a certain learning task). The randomization of participants to one of the two groups ensures that the characteristics of the participants are not biased in one of the two groups and guarantees that the error effects in an experimental studies are statistically independent (Recker, 2013). Pre-tests on certain characteristics help to determine whether the treatment and control groups have similar characteristics and ensure that the randomization was successful. In combination with post-tests, pre-tests are used to observe the impact of certain treatments on outcome variables.

Im my dissertation, I always conduct fully randomized experiments, meaning that a com-

putational mechanism randomly allocates participants to one of the conditions. Moreover, I usually control for different constructs with a pre-survey (e.g., as done in section 5). Next, I use post-surveys to capture certain behavioral constructs based on, for example, the technology acceptance model (Venkatesh et al., 2003; Venkatesh and Bala, 2008) or based on self-regulated learning (Bandura, 1991). These behavioral constructs ensure that the IT artifacts are appropriately designed (e.g., by measuring perceived ease of use as done in section 6) or help to explain the results of behavioral outcomes based on theories (e.g., as done in section 7). In the different experimental studies of my work, I try to provide the reader with a concise picture about the experimental task, the measured constructs, and the measured results by dividing the studies into *"pre, during, post"* treatment.

After conducting an experimental study, I apply a variety of statistical methods to analyze the impact of certain variables. For example, in section 6, I compare two different groups by conducting a t-test to access whether conversational argumentation tutoring has a significant effect on students' argumentation skills. The t-test helps to examine the differences in mean values between two groups. It can be used, for example, to determine whether the means of two sets of data are significantly different from each other.

Moreover, I apply analysis of variances (ANOVA) and post-hoc tests to measure the differences between two or more conditions. For example, in section 5, I present the results of a field experiment. Three different groups received different support when writing persuasive peer reviews over three month. To investigate if individual AI-based writing support impacts students' argumentation skills compared to general support, I use an ANOVA analysis and a Tukey post-hoc test to access whether the differences between the means between the groups is statistically significant.

## 3.5 Qualitative Methods

Besides using quantitative methods to investigate if and how AI-based argumentation learning support impacts students' learning process, I apply different qualitative methods especially for need-finding to ensure a user-centered design. In contrast to quantitative research methods, which are used to explain cause-and-effect dependencies, qualitative research is mostly based on inductive reasoning (Döring and Bortz, 2015). Inductive reasoning for qualitative studies is used to gain a deeper understanding of complex interdependencies in social or organizational contexts where human individuals interact. The paradigm of qualitative methods is based on situated activities that position the observer in the world. These methods usually involve a set of interpretive and material practices to ensure that the world is visible. The objective is to represent the real world in its

interdependence with data. This can include interviews, recordings, notes, or imagery (Recker, 2013). Hence, qualitative research methods are often based on more interpretive and naturalistic approaches to the real world. Qualitative researchers aim to study objects or subjects in their natural environment in order to better understand and interpret the phenomenon of interest (Döring and Bortz, 2015). These methods are especially of advantage when not much is known in advance about the phenomenon (Döring and Bortz, 2015).

Qualitative methodology is usually twofold: data collection and data analysis. Data collection includes methods such as interviews or observations (Recker, 2013). In this work, I use qualitative data collection through interviews in order to identify and summarize needs and user stories for the design of novel technology-mediated learning pedagogical scenarios. This helps me to systematically capture user needs before designing a new IT artifact. After the data is collected, the next step incorporates analyzing the results. Döring and Bortz (2015) name several concepts, such as coding, content analysis, or memoing as suitable methods of qualitative data analysis. In this work, I mainly use coding to develop underlying categories of the captured interviews as well as content analysis to semantically analyze the texts. In fact, I conduct user interviews to investigate how a certain user group is interacting or would like to interact with a certain technology in a certain scenario (e.g., an adaptive writing support system in a distance-learning scenario). The objective is to gain an initial understanding of the needs and requirements learners have for technology-mediated learning tools for argumentation. Therefore, I followed the expert interview method of Glaŝer and Laudel (2010), since the experts of their own usage are the users themselves. The interviews are guided through an interview guideline to ensure that all interviewees are confronted with the same questions. The user interviews are mainly used to derive user requirements for the design of novel educational artifacts. In order to more systematically collect user requirements, I not only code the interviews using qualitative content analysis but also derive user stories based on Cohn (2004). User stories usually consist of one or two written sentences and represent one desired functionality of the user. They usually follow a simple structure, for example:

*As *type of user*, I want to achieve *a goal* in order to *achieve something*.*

I particularly use the concept of semi-structured interviews and user stories in section 5 to inform the design and implementation of AI-based learning tools.

## 3.6 Design Science Research

The overall goal of my dissertation is to design and evaluate new forms of AI-based argumentation learning tools. In order to capture and guide the design research, I refer to the design science research paradigm (Hevner et al., 2004). Design science research is a common methodological principle in IS research that aims to guide the design and evaluation of novel IT artifacts to solve identified research problems (Hevner et al., 2004). The aim is to capture the design knowledge in a structured and reproducible fashion and evaluate it in order to form theoretical contributions (Gregor and Hevner, 2013). In the following paragraphs, I will provide the reader with a brief glimpse into the role of theory in design science research and then explain the main methodologies used in this work to generate novel design knowledge for AI-based argumentation learning systems.

### 3.6.1 Theory in Design Science Research

The design science research paradigm is based on the underlying principle that the construction of socio-technical artifacts themselves help to build knowledge and an understanding of design problems and their solutions (Hevner et al., 2004; Gregor and Hevner, 2013). In addition to truth seeking, design science research focuses on the utility of designed artifacts. These artifacts are the central component in design science. They can encompass man-made objects or processes such as decision support systems, modeling tools, governance strategies, or methods (Gregor and Hevner, 2013). The artifacts can be brought into material existence, e.g., into an abstract algorithms that can be instantiated into software (Gregor and Hevner, 2013). Gregor and Hevner (2013) provided a framework to position the knowledge contributions in design science. They distinguished between three different contribution types that differ in the level of abstraction, completeness, and maturity:

- **Level 1 contribution type:** situated implementation of artifact, such as instantiated software or implemented processes.
- **Level 2 contribution type:** nascent design theory knowledge as operational principles/ architecture, such as constructs, methods, models, design principles, technological rules.
- **Level 3 contribution type:** well-developed design theory about embedded phenomena, such as design theories (mid-range and grand theories).

In general, level 3 contributions (well-developed design theory) are more abstract, com-

plete, and mature knowledge compared to level 2 and level 1 contributions. Situated artifacts are, therefore, more specific, limited, and less mature knowledge. To better guide researchers in describing and communication their design findings as design theories, Jones and Gregor (2007) offer a framework with six core components and two additional components that a design theory should encompass. They identified eight distinct components of design theories: (1) purpose and scope, (2) constructs, (3) principles of form and function, (4) artifact mutability, (5) testable propositions, (6) justificatory knowledge (kernel theories), (7) principles of implementation, and (8) an expository instantiation. Table 7 illustrates the proposed framework of Jones and Gregor (2007).

| **Component** | **Description** |
|---|---|
| **Core Components** | |
| 1) Purpose and scope (the cause finalis) | "What the system is for," the set of meta- requirements of goals that specifies the type of artifact the theory applies to and in conjunction also defines the scope, or boundaries, of the theory. |
| 2) Constructs (the causa materialis) | Representations of the entities of interest in the theory. |
| 3) Principle of form and function (the cause formalis) | The abstract "blueprint" or architecture that describes an IS artifact, either product or method/intervention. |
| 4) Artifact mutability | The changes in state of the artifact anticipated in the theory, i.e., what degree of artifact change is encompassed by the theory. |
| 5) Testable propositions | Truth statements about the design theory. |
| 6) Justificatory knowledge | The underlying knowledge or theory from the natural or social or design sciences that gives a basis and explanation for the design (kernel theories). |
| **Additional components** | |
| 7) Principles of implementation (the causa efficiens) | A description of processes of implementing the theory (either product or method) in specific contexts. |
| 8) Expository instantiation | A physical implementation of the artifact that can assist in representing the theory both as an expository device and for purposes of testing. |

Table 7: Overview on the eight components of an IS design theory.
Source: Jones and Gregor (2007)

The components provide guidance to communicate and position the findings of design science research studies. They provide an overview of the most important dimensions of a design theory. The last two components are additional and not necessary, according to Jones and Gregor (2007). However, they are often explicated in a design theory. Gregor and Hevner (2013) also provide a framework to classify and structure the knowledge contribution of the design theory. Accordingly, the knowledge contribution is structured by the application domain maturity and solution maturity (see Figure 21). Most design sci-

ence projects are categorized by a high solution and application maturity. These research projects are routine designs without significant design theory contributions. For the other contribution types, Gregor and Hevner (2013) specify opportunities and contributions for design research.

In this dissertation, I focus on knowledge contributions in the domain of enhancement, i.e., analyzing and developing new solutions (AI-based argumentation writing support systems and pedagogical conversational agents) to known problems (adaptive support for students in large-scale learning environments).

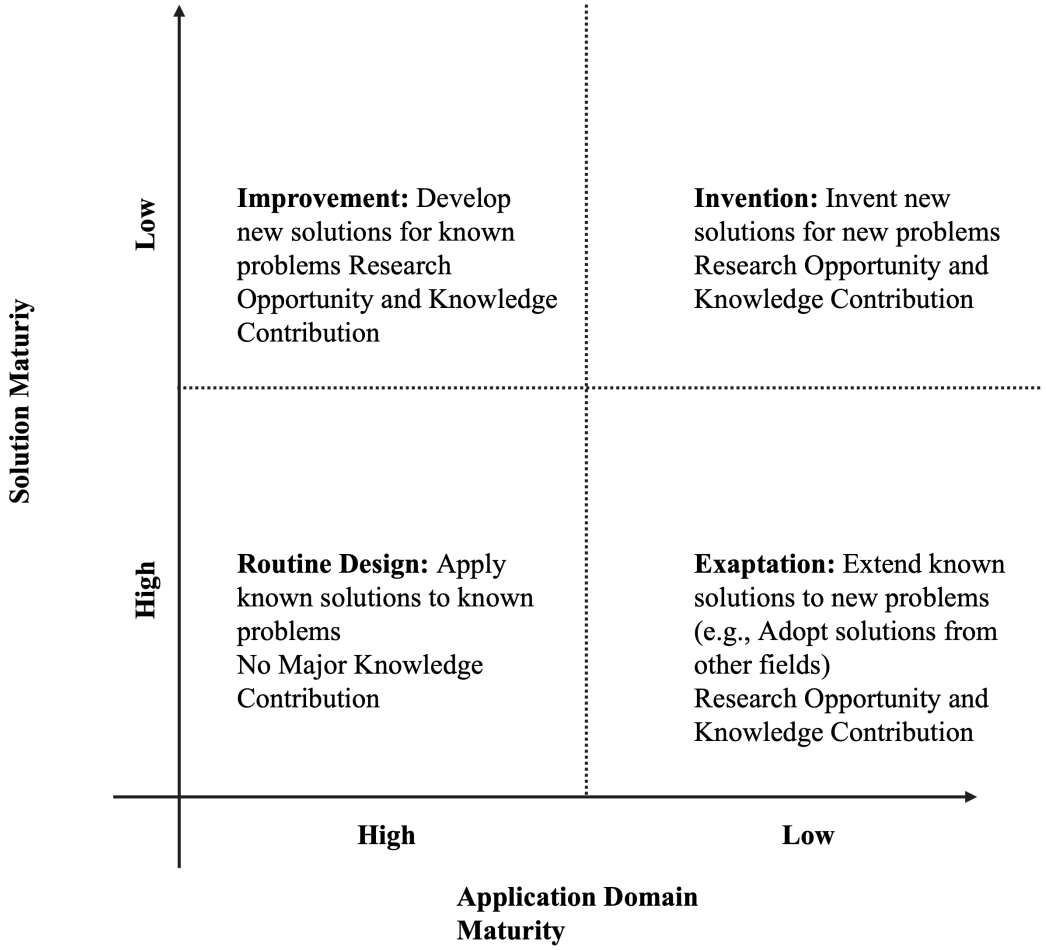



Figure 21: Classification of design science knowledge contribution types by the solution maturity and the application domain maturity.
Source: Gregor and Hevner (2013)

The formulation of design principles helps to specify design knowledge in an accessible and abstract form (Gregor et al., 2020). Several IS researchers have investigated what the best form of communicated general design guidelines, design patterns or design principles is (e.g., Gregor and Hevner (2013); Chandra et al. (2015); Gregor et al. (2020)). Gregor et al. (2020) proposed a conceptual framework that helps researchers to better specify the

different stakeholders when formulating principles (e.g., the end user, the designer, or the implementer; see Table 8). Hence, in this dissertation, I mostly follow the framework of Gregor et al. (2020) to position and to communicate my generalizable design knowledge in the form of design principles.

| **Design Principles XY** | |
|---|---|
| **Structure** | **Components** |
| *For Implementer **I** to achieve or allow Aim **A** for User **U*** | Aim, Implementer, and User |
| *In Context **C*** | Context<br>(Boundary conditions, implementation setting, further user characteristics) |
| *Employ Mechanisms **M1, M2, M3** ... Involving Enactors **E1, E2, E3**, ...* | Mechanisms<br>(acts, activities, processes, form/architecture, manipulation of other artifacts) Subsidiary components/artifacts that can have their own design principles |
| *Because of Rationale **R*** | Rationale<br>Theoretical or empirical justification for the design principle |

Table 8: The components of design principles to structure generalizable design guidelines. Source: Gregor et al. (2020)

In this work, I investigate different design principles for the design and development of AI-based argumentation writing support systems and conversational argumentation learning systems, for example in section 5. Hereby, I try to ensure that the design and the architecture of the developed learning tools are transferable to other domains and scenarios. In this vein, the addressed implementers (I) are usually educational designers. The aim (A) is, for instance, to enable students (user U) to receive adaptive feedback on their argumentation skills independent of an instructor, time, and location. I will specifically refer again to the framework of Gregor et al. (2020) when formulating the design principles in the different sections.

To form a transparent and reproducible scientific contribution to the knowledge base, different scholars have proposed frameworks to guide the often fuzzy and iterative design science research process (e.g., Peffers et al. (2007); Hevner (2007)). The next section will provide an overview of the most common design science process models.

### 3.6.2 Process Models to Guide Design Science Research

As Hevner (2007) states, *"it is vital that we [...] provide clear and consistent definitions, ontologies, boundaries, guidelines, and deliverables for the design and execution of high*

*quality design science research projects. Understanding and communicating the design science research process is essential not only to support acceptance among IS professionals but also to establish the credibility of IS design science among the larger body of design science researchers in the various engineering fields, architecture, the arts, and other design-oriented communities"* ((Hevner, 2007), p.1).

Several scholars have suggested frameworks to guide the iterative and often complex research process of design science (i.e., Hevner (2007); Hevner et al. (2004); Gregor and Hevner (2013); Peffers et al. (2007)). In this work, I mostly relied on two different frameworks, the design science process view according to Peffers et al. (2007) and the three cycle view of Hevner (2007). In the following, I will briefly introduce both methodologies.

The design science research process of Peffers et al. (2007) is based on six major steps that guide the research endeavor (see illustration of the process in Figure 22). The process starts with the problem identification and motivation of the research challenge and the value justification of the artifact. The second step incorporates the definition of objectives for a solution. This is often based on requirements for the to be developed artifact alongside quantitative or qualitative variables to define the success of the artifact for the prior stated problem definition. Next, the third phase deals with the design and development of the artifact, whereas the fourth step covers the demonstration of the artifact in a particular context. This can incorporate a suitable experimental study or a field instantiation. The fifth and the sixth step are concerned with the evaluation and the communication of the findings. Peffers et al. (2007) also state that their process view incorporates the iterative nature of design research. The process perspective, therefore, defines different entry points for the research process as well as process iterations to return to the second or third step after the evaluation to redefine the objectives for the solution or the design and development of the artifact.

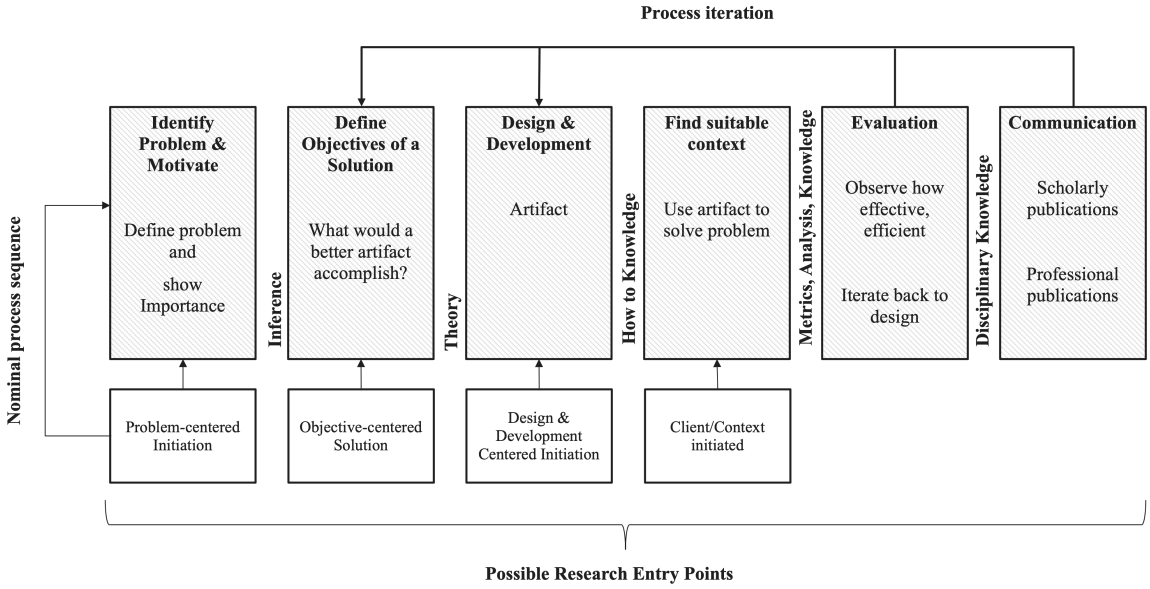



Figure 22: Process perspective of desgin science research according to Peffers et al. (2007).
Source: Peffers et al. (2007)

Nevertheless, in this dissertation, I mostly focus on the design science research methodology based on Hevner (2007). His three-cycle view emphasizes the interplay of relevance, design, and rigor cycles and, thus, focuses more on the iterative nature of the development of IT-based artifacts.

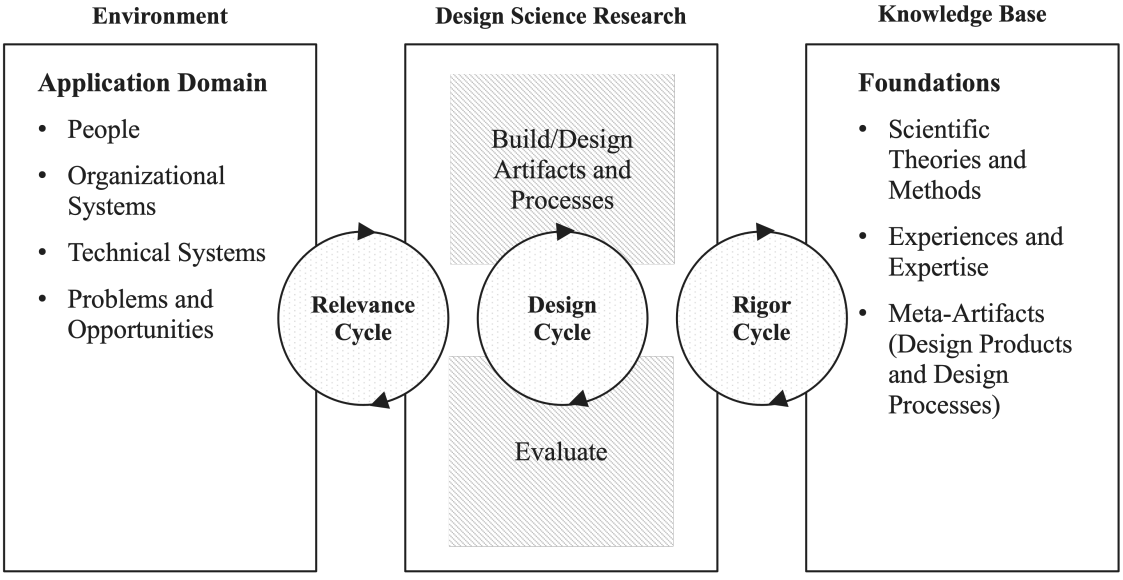



Figure 23: Overview of the three-cycle view to guide the iterative process of design science research.
Source: Hevner (2007)

The methodology is based on three different cycles, which are relevant for the successful application of a design science project (see illustration in Figure 23). The relevance cycle emphasizes the field requirements for the practical success of a design science project. It aims to bring the perspective of the corresponding real-world problem from the environment and the application domain of the artifact. In contrast, the rigor cycle refers to the role of theories and methods during the design process. It aims to connect the experience and expertise from the knowledge base with the design project and also emphasizes that the design science research project should add new knowledge generated by the research to the growing knowledge base (Hevner, 2007). In the center of the three-cycle view is the design cycle. Here, the focus is on the design, development, and evaluation of design artifacts and processes. The different cycles are connected and inform each other throughout the design science project. In the design cycles, for example, the requirements from all relevant stakeholders should be considered and iteratively tested with instantiated artifacts in real-world environments. The findings from the iterative evaluations can then be translated again into design knowledge (e.g., in design principles) and inform the knowledge base (Hevner, 2007).

The design evaluation plays an important role in design science research, as it provides feedback for further development and assures the rigorousness of the research (Venable et al., 2016). Different scholars have, thus, provided insights and suggestions to guide the evaluation of design evaluation (e.g., Venable et al. (2012, 2016); Sonnenberg and vom Brocke (2012); Nunamaker et al. (2015)). The "Framework for Evaluation in Design Science" (FEDS) provides guidance in developing a strategy for evaluating design science artifacts (Venable et al., 2016). The FEDS approach covers "why, when, how, and what to evaluate" (Venable et al. (2016), p. 1). FEDS considers a two-dimensional characterization of a design science evaluation iteration. One dimension deals with the functional purpose of the evaluation (formative or summative) and the other dimension covers the paradigm of the evaluation (artificial or naturalistic). Moreover, with their framework Venable et al. (2016) also offer a fourfold design process, which consists of (1) the goals of the evaluation, (2) the choice of the evaluation strategy, (3) the definition of the properties to evaluate, and (4) the design of the evaluation iteration.

In this work, I mostly organize the evaluation of my IT artifacts using the FEDS framework (Venable et al., 2016). For example, in section 5, I perform, first, a formative artificial evaluation of an AI-based argumentation writing support system in a laboratory experiment. Afterwards, I build on the design findings from the first evaluation iteration and conduct a summative naturalistic evaluation of the AI-based argumentation writing support system in a field experiment in a real-world lecture over three months. The FEDS framework, as well as the three-cycle view, supports this design science project and ensures that the communication of the results are transferable and reproducible and form

generalizable, realistic, and precise research contributions. This enables me to summarize my findings as a nascent design theory at the end of my work according to Jones and Gregor (2007).

In order to make the actual development process as well as the training of the ML algorithm for my IT artifacts more transparent, I rely on the cross-industry standard process for data mining. The next section will shortly introduce the reader to the different phases of this methodology.

### 3.6.3 Cross-Industry Standard Process for Data Mining

In this section the basics of the cross-industry standard process for data mining (CRISP-DM) based on the European strategic program on research in information technology will be briefly summarized (Chapman et al., 2000). CRISP-DM is a framework consisting of six sequences of activities, which describe steps according to which a data mining problem can be structured. It is an iterative cycle rather than a one-way process (see Figure 24).

The CRISP-DM model usually starts with a real-world data-related problem, e.g. using data to support a process with further knowledge. Therefore, the initial phase, the business understanding, is used to convert the objectives into a data mining or machine learning problem definition, including requirements and how outcome performance metrics are designed.

Subsequently, the required data is collected according to the requirements defined prior. Next, the data preparation phase is about cleaning, preprocessing, and transforming the target data. Cleaning might compromise noise filtering, missing value treatments, general preprocessing, feature extraction, feature reduction, and transformation of data into an applicable format for the modeling phase. The data preparation phase is one of the most import phases as the performance of the modeling depends significantly on the quality of the provided data.

Afterwards, clustering or classification algorithms are applied to the preprocessed dataset with different specific parameters in order to gain an optimal performance. Then, the results of the model are evaluated according to the defined business performance metrics. Finally, the deployment phase focuses on the usability of the model's output and makes sure the findings are applicable to the actual problem.

Besides CRISP-DM, two other major frameworks exist in literature that structure the process for building machine learning systems: The knowledge discovery process by Fayyad et al. (1996) and SEMMA (Sample, Explore, Modify, Model, Assess) by SAS Institute

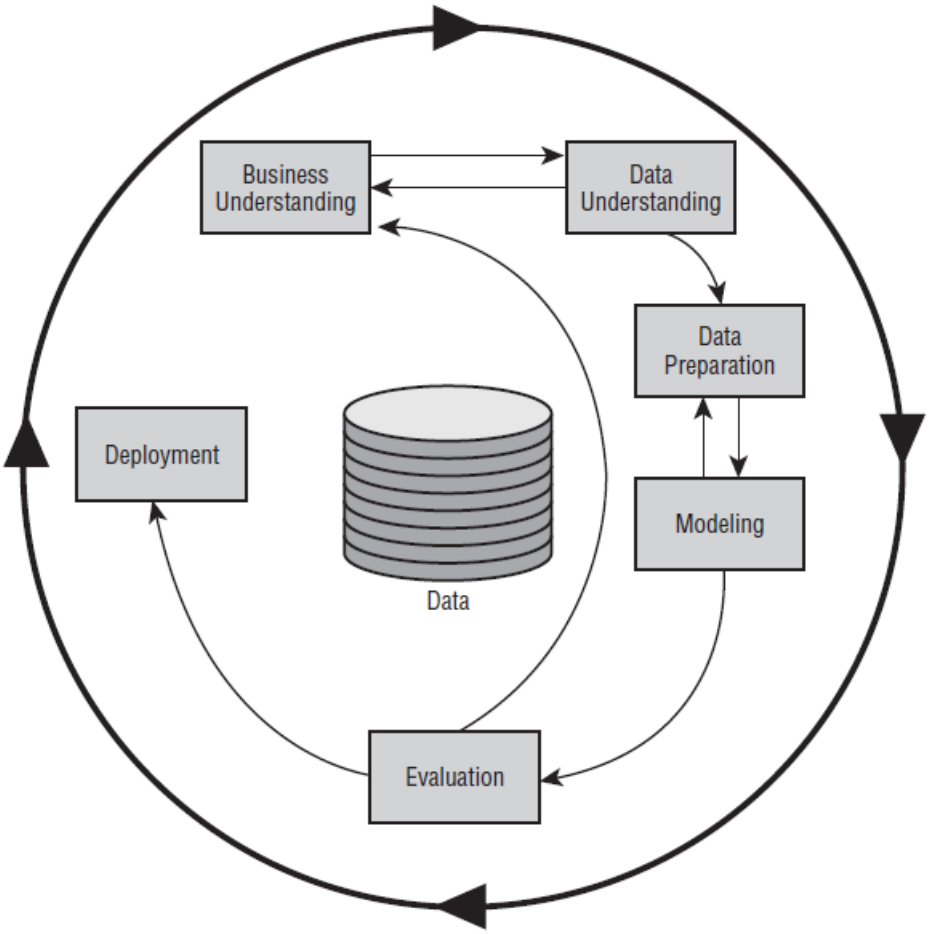



Figure 24: Structure of a data mining research iteration according to the CRISP-DM Model.
Source: Chapman et al. (2000)

Inc. (see Ramasubramanian and Singh (2017)). In this work, I rely on CRISP-DM, since it includes a problem-centered perspective and an iterative flow of knowledge between the different phases.

After explaining the underlying epistemological approaches as well as the research methodologies of my dissertation, the next sections will outline how and if AI-based argumentation learning systems help students to learn how to argue.

# 4 Developing an Understanding of Natural Language Processing for ML-based Information Systems

In this section of my dissertation, I will provide a structuration of the dimensions and characteristics of NLP for adaptive ML-based IS. This addresses and answers the first research question of my dissertation and will contribute to the technical foundation of AI-based argumentation learning tools in the course of my research:

**RQ1a:** *What are the theoretically grounded and empirically validated dimensions and characteristics of NLP features for ML-based IS?*

**Method:** *Systematic literature review, taxonomy development, and technical experiments.*

**Results:** *A taxonomy of NLP features for ML-based IS, structuration of usage clusters, transfer learning approach for argumentation identification cross text domains.*

Answering this research question paves the foundation of my dissertation since I provide a practical overview in the form of a taxonomy of NLP features for ML-based IS, such as AI-based argumentation writing support systems or conversational argumentation learning systems. Past structurations of NLP features for adaptive ML-based systems mostly lack a holistic design-driven perspective. Hence, I firstly aim to provide an overview of the literature of NLP features for ML-based IS in subsection 4.1. I conduct a systematic literature review based on Webster and Watson (2002) and vom Brocke et al. (2015) to investigate the manifold approaches for text feature engineering. Based on that, I developed a novel text feature taxonomy that aims to assist researchers and practitioners to classify, compare, and evaluate their NLP studies. Using the developed framework, a researcher can classify and compare a text feature according to its *linguistic analysis level, granularity level, information source, dimensionality*, and its *representation*. Moreover, I identified several application patterns and white spots in a cluster heat mapping analysis.

In the second part of this section, I dive deeper into on application area of this taxonomy: argumentation mining. By the time of starting this thesis, I developed the first approach towards domain agnostic argumentation identification with transformer models utilizing BERT Devlin et al. (2018) (see subsection 4.2). Both sections provide a foundation and framework of the technical development of the AI-based argumentation learning systems implemented in this dissertation. Such an understanding is necessary to concentrate the further course of this dissertation on the most important parts of my research.[3]

[3]The insights presented in this chapter are partly based on various papers (Fromm et al.,

## 4.1 Taxonomy of Features for NLP-based Machine Learning

### 4.1.1 Introduction

NLP is receiving continued attention from both researchers and practitioners. In the last years, several application areas have become popular, such as spam detection (Wood, 2016), sport performance prediction (Gruettner et al., 2020), web mining to predict pandemic outbreaks (Jahanbin and Rahmanian, 2020), or predictive policing (Meijer and Wessels, 2019). Devlin et al. (2018) have been very successful in processing large amounts of data and predicting outcomes precisely (Alom et al., 2019). For certain use cases, however, traditional ML models are often the more promising modeling approach, since only little labeled data is available, and transparency, as well as explainability, play a major role in the practical success of the application of those algorithms. Exemplary use cases can be found in human-centered decision support systems (Stoica et al., 2017; Alom et al., 2019), e.g., writing support systems for educational feedback (see section 5) or auditing for predictive policing (Scanlan, 2019).

The main challenge of NLP is to preprocess written text and extract valuable features to enable information extraction, data analytics, or ML algorithms to reach their maximum performance (Rajman and Vesely, 2004; Johnson et al., 2015; Allahyari et al., 2017). However, a large amount of NLP applications simply rely on basic feature generation techniques such as bags of words (Khadjeh Nassirtoussi et al., 2014) in which information such as the order and co-occurrence of words is not taken into account. This may lead to an overall underperformance, since copious training data is often not available (Pustejovsky and Stubbs, 2013). As a result, different teams of authors have indicated that enhancing NLP algorithms with the appropriate design, implementation, and evaluation of text features - a process commonly referred to as feature engineering - bears high chances of improving NLP outcomes significantly (Bird et al., 2009; Khadjeh Nassirtoussi et al., 2014; Johnson et al., 2015).

However, NLP feature engineering comes with a number of challenges. First and foremost, feature engineering still depends largely on human intuition and experience, since it requires deep domain knowledge to identify and operationalize relevant features (Ribeiro et al., 2020). Second, it is usually very time-consuming for specific or rare uses cases. In fact, according to surveys, data scientists spend about 80% of their time with data preparation, such as feature engineering (Anaconda, 2020; Forbes, 2017). Third, as Talib et al. (2016) point out, NLP has been influenced by different disciplines such as *Com-*

2019; Wambsganss et al., 2021a, 2020a). I thank my collaborators, the reviews, and the Associate Editors for the valuable feedback on my work.

*puter Science*, *Statistics*, *computer linguistics*, and IS. Accordingly, NLP features that have evolved in one discipline are often unknown or rarely used in other disciplines, e.g., in IS, human—computer interaction, or *computer linguistics*. This eventually leads to a fragmented body of literature, and sometimes contradictory research results concerning the impact of certain NLP features (Abbasi et al., 2018). An integrative view would be of utmost importance to systematically design, analyze, and compare the different configurations of text features for different use cases. A consistent knowledge base on the different characteristics and dimensions of NLP text features from a holistic perspective will help researchers and practitioners to systematically design, compare, and evaluate new or existing NLP applications. However, a systematic feature framework that helps researchers and practitioners to select from, compare, and evaluate new or existing features across different disciplines or areas of application is rather scarce. A few approaches that try to fill this gap already exist in literature, such as Ribeiro et al. (2020). However, these approaches usually focus more on the comparison of particular domain-specific features, e.g., through a checklist (Ribeiro et al., 2020), or on the learning algorithms but lack a holistic NLP feature classification framework. In this regard, IS research can offer a promising view to regard and classify a certain NLP-based ML use case from an interdisciplinary perspective into relevant elements.

Hence, my goal is to develop a comprehensive multidisciplinary taxonomy of text features to provide guidance by revealing the complex relationships of text features and by illustrating the application of certain feature groups based on similarity analysis. To achieve this goal, I (1) illustrate the diversity of text feature engineering through a systematic literature review in which I found 133 distinct NLP features in different applications domains, (2) outline a classification framework through a taxonomy with five dimensions and multiple characteristics, and (3) identify patterns and "white spots" in the application of text feature groups to provide further assistance for (a) the development and design of new NLP solutions and (b) the assessment and comparison of existing applications. For practical guidance, I, therefore, present the taxonomy as an NLP feature canvas in which a single text feature can be categorized and compared along five dimensions. Moreover, I illustrate all 133 identified text features on a web platform and encourage researchers to add missing features accordingly (see Wambsganss et al. (2021a)). The ultimate goal is to present an eclectic collection of text features and a corresponding classification framework towards a shared understanding across application domains.

The resulting taxonomy should simplify the comparability of text features between different studies, domains, or applications and reduce the costly effort of feature engineering for practitioners and scientists when traditional ML methods are applied. I present five dimensions of features diverse enough to demonstrate their commonalities and differences and comprehensive enough to encompass every feature I found in the systematic literature

review. I believe that such a menu of features will stimulate researchers and practitioners to enhance their knowledge about feature generation and to use the most appropriate features for their research.

### 4.1.2 Theoretical Background

#### 4.1.2.1 NLP Feature Engineering for Information Systems

Text-based AI for IS is based on techniques from different areas: NLP, data analytics, and ML. Hereby, the main challenge is to extract and reduce relevant features in a feature vector to reach a maximum performance by different ML algorithms (e.g., classification, clustering, or trend analysis, etc.) as depicted in Figure 25 (Rajman and Vesely, 2004). Features are understood as certain text characteristics or distinct attributes of a text that might bear valuable information for the ML algorithms. Thus, I use NLP and text features as synonyms.

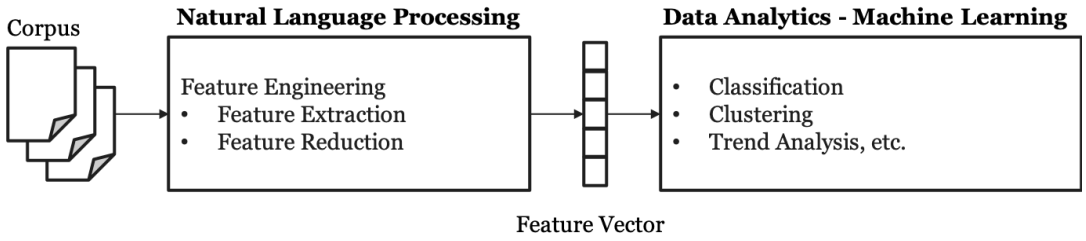



Figure 25: Illustration of pre-processing, feature engineering and NLP in an intelligent IS.
Source: Own illustration

Fueled with large amounts of data, neural networks, or transfer learning models such as BERT (Devlin et al., 2018) have been very successful in processing large amounts of data and achieving high-quality predictions (Alom et al., 2019). These models are said to learn features automatically in their top layers and adjust to the patterns of the data in their deep layers. However, big amounts of data are usually necessary for those algorithms to be trained to detect and predict these data patterns. Moreover, deep neural networks "do not reveal their features". This means that the user receives no indication of why the network has delivered a particular result. The lack of interpretability often hinders the practical application of neural networks. In fact, explainability has played a major role in the practical success of ML algorithms, as they are often embedded in human-centered decision support systems (Stoica et al., 2017; Alom et al., 2019). On the other hand, traditional ML approaches often show better performance for small amounts of input data (e.g., less than 1,000 labeled data points) (Alom et al., 2019). Traditional

ML models together with text feature engineering are therefore still mainly used as the default modeling approach for NLP, since often only small data sets are available that are labeled, e.g., for human-centered writing support systems (see section 5) or auditing for predictive policing (Scanlan, 2019).

Bird et al., (2009, p. 224) state that "selecting relevant features and deciding how to encode them for a learning method can have an enormous impact on the learning method's ability to extract a good model". Thus, most work in building a text classifier with traditional NLP methods is creating relevant features and deciding how to represent them. Bird et al., (2009, p. 224) mention that it is possible to receive a decent performance "by using a fairly simple and obvious set of features ...", however, "there are usually significant gains to be had by using carefully constructed features based on a thorough understanding of the task at hand". Nevertheless, most features are created through a process of trial and error and not by rigorous feature engineering supported by a framework or a taxonomy. In fact, I did not find a comprehensive and concise taxonomy for NLP features in the literature. What can be found are classifications of NLP features along single dimensions, which are often more a simple categorization of attributes than a set of comprehensive and robust dimensions (e.g., Johnson et al. (2015); Missen et al. (2013)). Mostly, feature engineering is done by intuition about what information might be relevant to the problem. As Bird et al., (2009, p. 224) wrote, "it's common to start with a *kitchen sink* approach, including all the features that you can think of, and then checking to see which features actually are helpful."

As a consequence, the different application areas of text analysis and NLP have created a variety of features. Education and literary sciences have brought up a number of readability indices, which are used to judge the readability level of books and texts. The one most widely applied is the Flesch readability index, which is calculated from the number of sentences, words, and syllables of a text (Flesch, 1943; Kincaid et al., 1975). Sentiment analysis and opinion mining are concerned with the polarity of texts. Polarity is a feature describing the emotion or sentiment present in a text. It can be observed on the word, the sentence, or the document level (Pang and Lee, 2008). Applications for authorship attribution and verification analyze the syntactic depth and complexity of texts and have come up with features like vocabulary richness of the use of hapax legomena (e.g., NagaPrasad et al. (2015)). Requirements analysis is concerned with the question if textual system requirements are clearly and concisely formulated. Accordingly, features are used that identify words, subjunctives, or passive voice, which are all indicators for ambiguous language. Researchers use argumentation mining to develop algorithms that extract argumentative components and discourse structures from given texts (Lippi and Torroni, 2015b). Stab and Gurevych (2017), for example, used, as they called it, a combination of structural, syntactical, lexico-syntactical, and probability features for identifying

argument components in persuasive student essays (Stab and Gurevych, 2017).

All these application areas and disciplines are confronted with different challenges and therefore have developed their own features and their own modes of speaking. A holistic feature classification framework might help by illustrating the features' commonalities and differences. However, current feature frameworks fall mostly short of depicting the nature of text features across the different disciplines.

#### 4.1.2.2 Taxonomies as a Means of Classifying NLP Features in IS

To better understand the relevance of a taxonomy for NLP features for research and practice, I must first understand what a classification is and what it may be used for in the context of (a) the development and design of new NLP-based ML use cases and (b) the assessment and comparison of existing NLP-based ML applications. Fundamentally, a classification of objects serves as a cognitive aid that enables researchers and practitioners to navigate complexity caused by several variables of interest (Hambrick, 1984; Nickerson et al., 2013). There are different ways to develop classifications of which taxonomies are the most famous ones in IS (Bailey, 1994). A taxonomy is a theoretical study of classifications and can include different objects such as principles, procedures, and rules (Bailey, 1994; Usman et al., 2017). Although initially taxonomies were developed according to an empirical base, nowadays, conceptual bases are additionally used (Nickerson et al., 2013). Thereby, taxonomies go beyond classifying objects as they ease knowledge sharing, provide a better understanding of interrelationships among objects, and thus support decision making (Bailey, 1994; Usman et al., 2017). Moreover, assuming an abductive reasoning approach, they may provide a very different understanding about how NLP features in IS may be used and, from a pragmatic perspective, about how these elements might support researchers and practitioners in developing NLP applications. Thus, taxonomies may change the way I think about NLP feature engineering in IS since they go beyond merely classifying objects and additionally serve as a way to predict outcomes (Doty and Glick, 1994). The development and design of new NLP solutions in IS requires knowledge on the anatomy of NLP features and key factors that need to be met to train, build, and deploy a new NLP application. In this sense, the ability to classify NLP text features is a prerequisite to building, comparing, and assessing NLP-based ML use cases. However, current literature falls rather short on a comprehensive and robust structuration of NLP features. Recently, Ribeiro et al. (2020) provided a tool called "checklist" that aims to assist NLP researchers in testing features "in the wild" by providing a sort of checklist tool, however, their research was not based on a systematic literature review and did not incorporate certain feature levels. Therefore, I aim to address this literature gap

by investigating a novel taxonomy that supports decision-making in building, designing, and comparing NLP features, and helps to specify the relationships of the very features towards the outcome of a trained model.

### 4.1.3 Methodology

The research started with a systematic literature review based on Webster and Watson (2002) and vom Brocke et al. (2015) to capture the diversity of text features used for feature engineering across different disciplines (see Table 9). Second, I derived dimensions and characteristics in a novel taxonomy following the method presented by Nickerson et al. (2013). Third, I evaluated the quality of the taxonomy with five semi-structured interviews with experts from research and industry. Finally, I performed a cluster analysis to display the complex relationships of text features in a heat map and to identify the application of certain feature groups that used similar features (Kaufman and Rousseeuw, 2005).

#### 4.1.3.1 Step 1: Database Creation Through a Systematic Literature Review

To identify the diversity of different text features used for feature engineering, I have drawn on the systematic literature search approaches by Webster and Watson (2002); vom Brocke et al. (2015). Based on current literature on feature engineering such as Bird et al. (2009); Weiss et al. (2010); Missen et al. (2013); Ribeiro et al. (2020); Johnson et al. (2015), I identified different keywords, which researchers used to describe the process of NLP feature engineering. Based on these, I build the following search string trying to incorporate the previously captured keywords: *("text OR language OR handcrafted OR manual OR human") AND ("Feature") AND ("Engineering OR Generation OR Selection OR Creation").* To find relevant literature with NLP studies that applied feature engineering, I applied the search string to the following six databases: AISeL, EBSCO, Science Direct, ProQuest ABI Inform, and ACM Digital Library. The papers were retrieved from January to May 2019.

**Paper selection:** The database search resulted in 9698 hits. Titles, abstracts, and keywords were screened to fit the above-mentioned definition and application to the scope of the study. I excluded papers that did not refer to textual data or that applied feature engineering in another domain than NLP. Multiple papers were excluded due to a different research scope described in their abstract, e.g., several papers described feature extraction

| | **Step 1: Database Creation** | **Step 2: Taxonomy Development** | **Step 3: Taxonomy Evaluation** | **4: Taxonomy Application** |
|---|---|---|---|---|
| **Steps** | Search for relevant papers in IS and NLP literature; analyze and synthesize literature concerning NLP feature engineering | Define meta characteristic; run taxonomy development iterations | Evaluate dimensions and characteristics with experts based on quality criteria | - Cluster analysis to display relationships of NLP in a heat map |
| **Method** | Literature Review (Brocke et al., 2015), Qualitative Coding | Taxonomy Development (Nickerson et al., 2013) | Expert evaluation (Szopinski et al., 2019) | Cluster analysis and heat mapping (Kaufman and Rousseeuw, 2005) |
| **Source** | NLP feature engineering literature | Existing classifications, database of 133 features | Semi-structured interviews with experts | NLP literature (identified in step 1) |
| **Result** | Database with 211 articles on NLP feature engineering | Taxonomy of NLP features | Evaluated taxonomy of NLP features | Insights of NLP usage patterns |

Table 9: Overview of the four consecutive research steps for deriving a taxonomy for NLP-based features in adaptive IS.
Source: Own illustration

for text recognition in images and thus were eliminated from the sample. This screening process resulted in 241 potentially relevant papers that mentioned that they applied feature engineering in the course of their study. After the elimination of all duplicates, 206 relevant papers were remaining. Afterwards, a forward and backward search was carried out according to Webster and Watson (2002). Through screening the references, 17 of the articles were added to the list, resulting in 223 relevant papers. Next, I thoroughly read each paper. I found 12 papers that described a process or applied a method within NLP without mentioning or describing any feature. Therefore, I excluded those papers from the sample and finally ended with 211 papers that discussed feature engineering for NLP and described at least one textual feature.

**Paper analysis:** The 211 relevant papers were analyzed from a concept-centric perspective based on an abductive approach. Thereby, to aggregate the insights from identified NLP studies, I developed a list of master codes and master code descriptions represent-

ing different NLP features. This process was iterative and required multiple rounds of coding of the identified papers by different researchers. The process started with two of the researchers independently coding a subset of 25 randomly chosen articles. For each of the 25 studies, we listed the used NLP text features and the application domain. If the features were not directly mentioned in the study, we tried to investigate the features in the appendices or online representations of the research. We conducted a workshop to discuss how to combine NLP features across studies, which resulted in a distinct list of text features and descriptions. During the next iterations, one researcher always coded a batch of 25 articles based on the feature list and definitions recorded so far. Afterwards, a group of three researchers met to discuss the findings. In case the coding was unclear, the NLP features, as well as the descriptions, were discussed and corrected until an agreement was achieved. In each iteration, we added new features and descriptions to the list until all papers were coded. The final feature list consisted of 133 unique NLP features from 211 coded papers.

#### 4.1.3.2 Step 2: Taxonomy Development

I aim to provide framing through the development of a comprehensive taxonomy. Therefore, I follow the method presented by Nickerson et al. (2013). All ending conditions Nickerson et al. (2013) named should be fulfilled by the final taxonomy (see section 3). In the presented taxonomy below, all ending conditions are met and I believe that to the best of my knowledge the final taxonomy represents all the found NLP features. However, especially condition D) might give me difficulty in the future since I would never claim that I found all NLP features used by practitioners or scientists. However, I believe that the current taxonomy presents the diversity of text features and completeness to all identified features in the systematic literature review. I have found a set of features diverse enough to demonstrate their commonalities and differences and comprehensive enough to display all identified text features.

Since I would estimate a high level of knowledge in the research area and multiple NLP studies with feature engineering are available, I conducted a conceptual-to-empirical cycle first, followed by four empirical-to-conceptual cycles. I inductively challenged the latest status of the taxonomy by classifying NLP features and revising existing dimensions and characteristics accordingly. In total, I classified all of the 133 found NLP features in five iterations until all ECs were met. Finally, the taxonomy consists of five dimensions and multiple associated characteristics.

#### 4.1.3.3 Step 3: Taxonomy Evaluation

To ensure the quality of the taxonomy, I assessed it against the following five criteria: conciseness, robustness, comprehensibility, extendibility, and explanatory by Nickerson et al. (2013). Hence, to evaluate the taxonomy, I conducted semi-structured interviews with five experts that either had expertise in NLP research (four interviews) or NLP development in practice (one interviewee) following the taxonomy evaluation suggestions of Szopinski et al. (2019). The shortest interview lasted 45 minutes, while the longest took 56 minutes. The interview guideline consisted of 18 open questions which were based on the five evaluation criteria. The final version of the taxonomy, the meta-characteristic, as well as exemplary NLP feature, were sent to the interviewees before the interviews. In the interviews, I then asked the interviewees to comment and identify potentials for revision and improvement of the taxonomy.

All in all, the interviewees were quite positive about the conciseness of the taxonomy. All of them mentioned that the number of dimensions was well chosen and do not overwhelm the reader. Concerning the robustness, the experts mentioned that some feature groups are only possible with certain characteristics, e.g., the semantic linguistic level is almost always a corpus-based feature, since external information is needed for the semantics. I considered this in the evaluation, however, all of the experts agreed some combinations are just natural in NLP text features and therefore also dependent on each other. Next, the experts agreed that the taxonomy is comprehensive concerning the state of the art. However, they also claimed that based on ongoing rapid developments in the area of NLP, new dimensions and characteristics may need to be added in the future. Nevertheless, they agreed that the taxonomy probably illustrates the state of the art NLP text features. In fact, all experts mentioned that the taxonomy is easily extendible, and novel features could easily be added later on. Concerning the explanatory power, the experts were convinced that the taxonomy is easily understandable for an experienced NLP researcher or practitioner, However, might be harder for a novice user to understand, since it demonstrates the complex feature relationships. I followed that up by illustrating the final taxonomy as an NLP feature canvas with guiding questions to also enable novice NLP practitioners to easily understand the complex relationship of NLP features (Figure 27).

#### 4.1.3.4 Step 4: Taxonomy Application through Cluster Heat Mapping

Based on the finally evaluated taxonomy, I aimed to provide further insight by displaying the complex relationships of text features and their application to provide researchers

and practitioners further support in their practical work. Text features are rarely used individually, but often in combination with other, domain-related features. Such groups of related features evolved in different disciplines and application areas. Therefore, I argue there may exist "natural" clusters of NLP features, which would help me to display the complex relationships of text features and better interpret their application to provide researchers and practitioners further support in their practical work. I used cluster analysis and heat mapping as a descriptive, exploratory tool to identify these natural patterns in data (Kaufman and Rousseeuw, 2005). For this purpose, I created a data matrix D(i,j), $1 \leq i \leq N$, $1 \leq j \leq M$ with the rows representing the N features and the columns representing the M publications. A matrix element D(i,j) is set to 1, if publication j contains feature i, and 0, if not. This resulted in a binary NxM data matrix. Since the individual features were created from the systematic literature review, they are already pre-ordered according to their granularity-level: morphological, word, sentence, document, and beyond-document. For the publications in the columns, there is no such "natural" order. They are simply organized in alphabetical order.

I performed a clustering based on Ward's algorithm (Ward 1963) as an agglomerative clustering technique with a Euclidean metric since it is known for accurate clustering with smaller data sets. I used the Python-based API SciPy as well as Matplotlib for plotting the resulting matrix in a heat map. The result of hierarchical clustering is a total ordering of the objects in form of a tree or dendrogram, the leaves of which define a similarity sequence. This sequence can be used to rearrange the columns of the matrix in such a way that columns that are most similar to each other become directly adjacent. In other words: publications that use similar sets of features are grouped closely together. I believe, by visualizing the results, I can provide further insights by identifying a natural grouping of publications that used similar features and derive possible white spots.

### 4.1.4 Taxonomy of NLP Features for ML-based IS

In total, I found 133 distinct NLP features in 211 papers that discussed feature engineering for NLP and described at least one textual feature. I found three groups of papers: (1) papers that described features but did not give them a designation (= a class name); (2) Papers that used such designations for features, which I considered as the starting point of a classification; and (3) a group of papers that presented lists of features and classified them along a single dimension and, in very rare cases, along two dimensions (e.g., Abbasi et al. (2008)).

20 papers contained a feature list or table and 191 papers simply present their distinct features without further illustration or categorization (see Figure 26 on the right). The

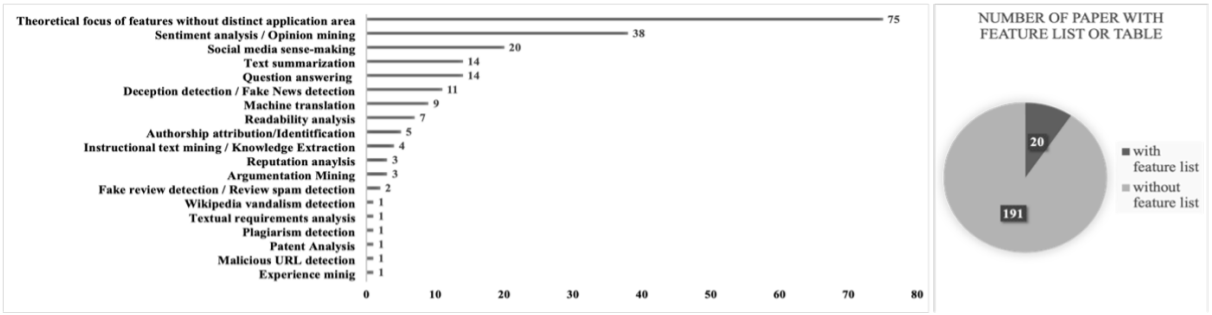



Figure 26: Overview on analyzed NLP features across domains.
Source: Own illustration

majority of papers that I analyzed applied feature engineering to a specific NLP domain. 75 of the 191 papers had a theoretical focus without addressing one particular application area. The remaining 136 papers, however, could well be assigned to a particular application area. Figure 26 shows their distribution on the left. In total, I found more than 20 application areas in the sample of NLP papers. Most of the publications were applying feature engineering to sentiment analysis (20 percent) and social media sense-making (10 percent). The rest of the sample applied NLP feature engineering to various other topics, such as authorship attribution, deception detection (e.g., fake news detection), text summarization, question answering, readability analysis, or argumentation mining. Over 90 percent of the studies conducted their NLP research on English language texts. Other studies included German, Japanese, Spanish, French, or Mandarin documents.

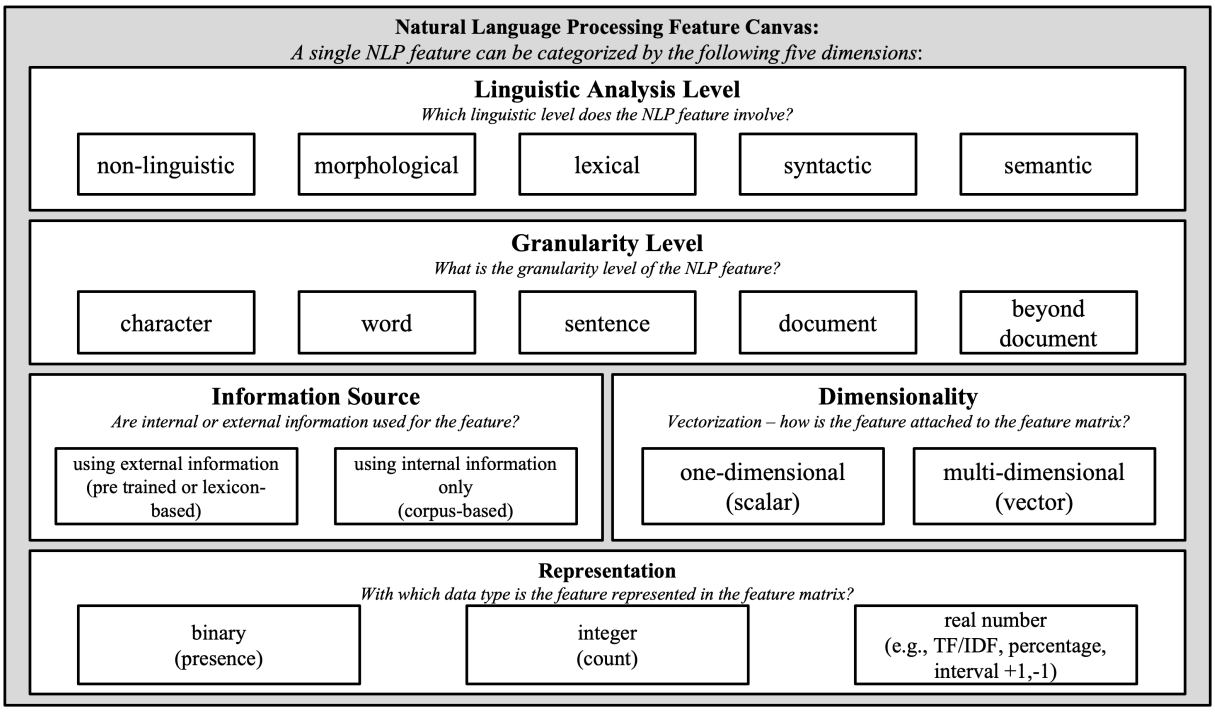



Figure 27: Taxonomy of NLP features in an NLP feature canvas.
Source: Own illustration

#### 4.1.4.1 Linguistic Analysis Level

The dimensions most widely mentioned in literature are the *linguistic perspectives* (Johnson et al., 2015), or in other words, the hierarchy of stages in NLP: morphological analysis, lexical analysis, syntactical analysis, and semantic analysis (Johnson et al., 2015; Bird et al., 2009). Lexical analysis converts a textual input stream into words (also called terms or tokens) that build the vocabulary of the text. Morphological analysis looks at the internal structure of the words in more detail (e.g., syllables). Syntactic analysis determines the structure of a sentence and the role that each word has within the sentence. Semantic analysis is concerned with the meaning of a word, a sentence, or a whole text. Many authors classify NLP features along with these linguistic perspectives. Lexical, syntactic, and semantic features are described by Kambhatla (2004) for relation detection, by Abbasi et al. (2008) for sentiment analysis, by Loni et al. (2011) for question classification, and by Alzahrani et al. (2012) for plagiarism detection. Hancke et al. (2012) use morphological, lexical, and semantic features for readability classification. NagaPrasad et al. (2015) distinguish lexical, semantic, and structural features for authorship attribution. Van Den Bosch (2017) describe lexical and syntactic features for language variety identification. The problem that linguistic analysis levels create when developing a feature taxonomy is that there is not necessarily a one-to-one relationship between a feature and an analysis level. Sentiment analysis works basically with the meaning of words (with positive or negative sentiments) and thus requires a lexical and semantic but not necessarily syntactic analysis (Hatzivassiloglou and Wiebe, 2000). The study of meanings of words is called lexical semantics (Johnson, 2007). In the analysis, I always assign a feature to the highest linguistic analysis level. However, there are also features that are not presented through a linguistic category, such as the number of replies to a post (Abbasi et al., 2008). Examples are also stylistic features (Cossu et al., 2015; Mahajan and Zaveri, 2017), contextual/ non-contextual features (Negi and Buitelaar, 2014), and frequency-related and intensity-related features (Mahajan and Zaveri, 2017). I addressed those features with a category called *nonlinguistic features*.

#### 4.1.4.2 Granularity Level

A much less observed characteristic of textual features is their granularity level (called in Missen et al. (2013) and Suh (2016)). A hierarchy can be described that decomposes a text (a document, a review, a post) into sentences, which are decomposed into words, which are decomposed into characters. Even if the granularity levels seem to be similar to the linguistic analysis levels, they are not the same. This becomes vividly illustrated by

| **Feature** | **References** | **Linguistic Analysis** | **Granularity** | **Ext. Source** | **Dimensionality** | **Representation** |
|---|---|---|---|---|---|---|
| *Frequency of character "@"* | (Zheng et al., 2006) | lexical | character | no | scalar | integer (count) |
| *Average number of syllables per word* | (Feng et al., 2010) | morphol. | word | no | scalar | real number |
| *Average sentence length in words* | (Suh, 2016) | lexical | word | no | scalar | real number |
| *POS-n-gram vectors* | (Brett and Pinna, 2015) | syntactical | word | no | vector | integer (count) |
| *Word embeddings* | (Levy and Goldberg, 2014) | lexical | word | yes | vector | real number |
| *Binary character trigram vectors* | (Lipka and Stein, 2010) | lexical | character | no | vector | binary (presence) |
| *Readability indices (Flesch, etc.)* | (Flesch, 1943) | lexical | document | no | scalar | real number |
| *Word polarity* | (Agarwal and Mittal, 2016) | semantic | word | yes | vector | real number (-1;+1) |
| *Sentence polarity* | (Missen et al., 2013) | semantic | sentence | yes | scalar | real number (-1;+1) |
| *Document polarity (Review polarity)* | (Mukherjee and Bhattacharyya, 2012) | semantic | document | yes | vector | real number (-1;+1) |
| *Time stamp of a message* | (Abbasi et al., 2008) | nonlinguistic | document | no | scalar | real number |
| *Title similarity across documents* | (John et al., 2017) | lexical | beyond document | no | vector | real number (TF/IDF) |
| *Number of links pointing to a web page* | (Fürnkranz, 1999) | nonlinguistic | beyond document | no | scalar | real number |

Table 10: Feature examples categorized by linguistic analysis level, granularity level, external sources, dimensionality and representation.
Source: Own illustration

looking at unigrams and POS unigrams (Brett and Pinna, 2015; Reyes and Rosso, 2012). A text *"the highest peak in the country"* (Brett and Pinna, 2015) is composed of the single words (= unigrams) *"the"*, *"highest"*, *"peak"*, *"in"*, *"the"*, *"country"* – which have the syntactic roles (= part-of-speech tags, POS tags) *"AT0"*, *"AJS", "NN1", "PRP", "AT0", "NN1"*. Occurrences of both unigrams and POS unigrams can be counted and

thus be used as features. It is clear that unigrams require lexical analysis and POS unigrams require syntactic analysis (*"part-of-speech tagging"*). So, unigrams and POS unigrams are on different linguistic analysis levels, but they have the same granularity; they are both on the word level. Thus, the granularity level *"word"* does not necessarily coincide with the linguistic analysis level *"lexical"*. Another example is the semantic feature *"polarity"* (semantic orientation). Polarity can be analyzed on the word level, on the sentence level, or on the document level (Missen et al., 2013), which again shows how important the granularity level is for differentiation. On top of that, I realized that several authors used features that go beyond the level of the single document, e.g., the similarity of titles between multiple documents (John et al., 2017). Here, lexical components are compared between different data points (documents). Moreover, there exist plenty of nonlinguistic features that go beyond pure text analysis. Examples are the number of links pointing to a web page (Fürnkranz, 1999), the number of clicks on a question and answer pair (Jeon et al., 2006), or the number of tIets marked as favorites (Cossu et al., 2015). This type of feature depicts a characteristic that is above the single document level, mostly looking at a collection of several documents. This can be hypertexts, discussion threads on web forums, or question and answer threads. I address the granularity level of those features with a characteristic called "beyond document".

#### 4.1.4.3 Information Source

Semantic features often rely on externally available lexicons that describe the semantic orientation or polarity (positive or negative) and subjectivity (subjective or objective) of individual words or phrases (Taboada et al., 2011). SentiWordNet is such a lexical resource used for opinion mining (Baccianella et al., 2010). Since semantic features can also be corpus-based (Liao and Grishman, 2010), a dimension describing the use of external sources is appropriate. Therefore, I included a dimension called "information source" with the characteristics corpus-based and lexicon-based to provide differentiation.

#### 4.1.4.4 Dimensionality of Features

As already stated above, several classifications of NLP features along single dimensions can be found in the literature. A dimension that has rarely been brought up in research papers is the dimensionality of a feature – probably because it is too obvious for most of the authors. There are essentially two cases: features that are expressed in a single number, like the number of occurrences of a specific word in a document, the percentage of nouns in a text, or the Flesch readability index (e.g., Feng et al. (2010); Kincaid et al.

(1975); Flesch (1943)), and features represented as a vector, like bags of words, bags of n grams, word embeddings, or sentence embeddings (e.g., Brett and Pinna (2015); Wambsganss et al. (2020a); Levy and Goldberg (2014)). The dimensionalities of the latter correspond with the size of the individual vocabulary or embedding dimensions. Therefore, it is sufficient to distinguish the two characteristics as one-dimensional (scalar) and multi-dimensional features (vector).

#### 4.1.4.5 Representation of Features

All features that I found in the literature were expressed in numbers. They can be distinguished according to their representation (Khadjeh Nassirtoussi et al., 2014) in binary numbers (0, 1), integer numbers (e.g., counts, frequencies), and real numbers (e.g., percentages, values within an interval). Bags of words appear in four different representations: presence of a word in a document (binary), number of occurrences of a word in a document (TF = term frequency), TFIDF representation of a word in a document (TFIDF = term frequency-inverse document frequency), and a real number representing the relative importance of a word in a document within a given corpus. Since TFIDF has a paramount role in NLP, I keep TFIDF as a separate class within the dimension *"representation"*. Literature often does not distinguish clearly between the name of a feature and its representation. Some authors call an individual word a feature, some say that the presence or the number of occurrences of a certain word is a feature. Günal et al. (2006) talk about 140 features for spam detection, and they present a list of 140 words (Günal et al., 2006). With the representation dimension, things should become clear. I want to provide a deeper understanding of the five dimension and their characteristics. Therefore, I allocated different feature examples in the derived dimensions illustrated in Table 10.

### 4.1.5 Cluster Heatmapping of NLP Features Across Studies

To address the third research goal, I aim to identify patterns of text features by illustrating the frequency of text features according to the dimensions and characteristics I found in the sample of studies (see Figure 28). Moreover, I unveil complex relationships between text features by performing a hierarchical cluster analysis on the studies and by visualizing the results in a heat map (see Figure 29). In Figure 28, I illustrate the frequency distribution of NLP features I found in the subset of NLP studies. With respect to the linguistic analysis level, 74 percent of all features are on the lexical level. Moreover, 73 percent of studies used features on the word granularity level, and around 27 percent

used external information such as dictionaries for feature engineering. The analysis shows that researchers solely use features above the word or lexical level.

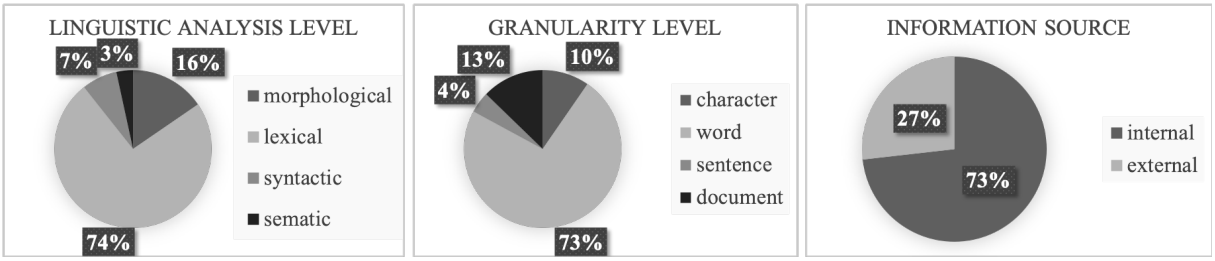



Figure 28: Overview on the sample of 211 NLP research, the frequency distribution of text features for the linguistic analysis level, granularity level, and information source.
Source: Own illustration

Also, the objective was to identify patterns and white spots in order to better interpret the application of certain text feature groups which used similar features. I do not aim to derive a finite set of clusters with defined propositions. In fact, the objective is to contribute to literature and practice by deepening the understanding of the used features and describing the distribution of NLP features of the sample of studies. In Figure 28, I display the binary *NxM data matrix* after rearranging the columns based on *Ward's algorithm* (Ward, 1963). The *rows* represent the *N NLP* features (N=133), and the *columns* encompass all *M studies* from the sample collected in the systematic literature review (M=211). A black dot indicates that a feature is used by the corresponding study. On the left side, I illustrated the resulting tree in a dendrogram, the leaves of which define a similarity sequence.

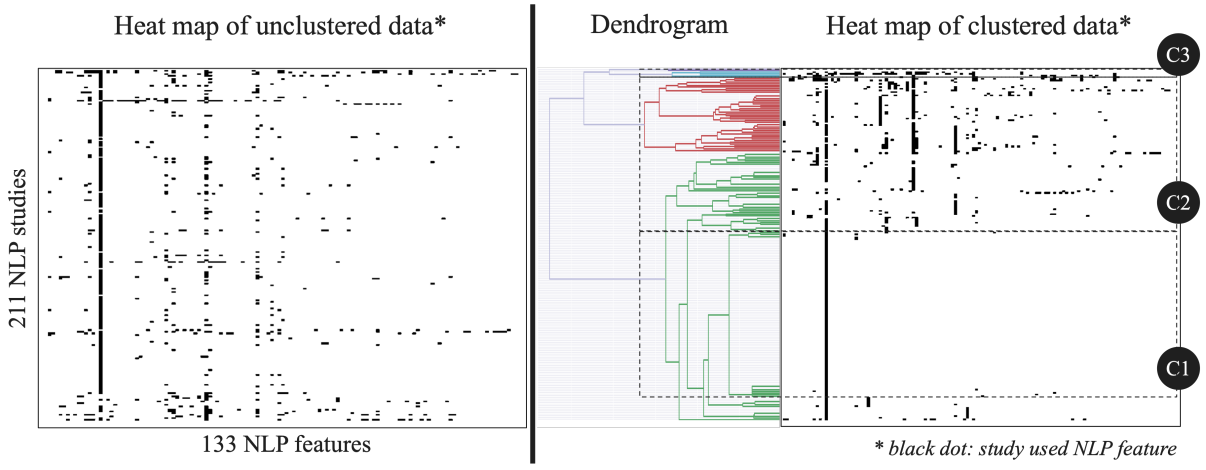



Figure 29: Illustration of the unclustered and clustered data matrix in a heat map. Lines of the matrix represent publications, columns represent features.
Source: Own illustration

In the heat map, one can see filaments in the data, which indicate a certain feature and a group of studies that use it. Most prominent is the lexical word-level feature bags of

words (BoW). More than 90 percent of the analyzed papers were using bags of words, which explains the long black vertical line in the heat map. Based on the dendrogram, I identified three main clusters of application patterns of studies (C1, C2, C3). Cluster one (C1 in Figure 29) represents a group of studies that solely used bags of words as a feature. In fact, 50 percent of the NLP applications in the data set are based on this word-level feature, in which the order and co-occurrence of words are not considered. Either the authors were not aware of higher-level features or they were satisfied with the performance of their analysis based on bags of words. However, since many further features are available and could be of value, these studies in C1 would have the potential to improve their results by conducting feature engineering, e.g., based on the taxonomy. In cluster two (C2), one sees a group of studies that use bags of words in combination with other features (around 40 percent). These features are indicated by two more filaments: One feature group is POS grams, which 32 percent of the analyzed papers used. Another band reflects the occurrence of negative or positive words, which 22 percent of the studies in the sample used. Finally, I found a third cluster (C3), which reflects a small group of papers which used intensive feature engineering across all granularity- and linguistic-levels in their NLP-studies. However, these studies reflect about five percent of the overall papers in the data set. This indicates that there seems to be untapped potential in current NLP studies.

### 4.1.6 Discussion

With this research, I analyzed the manifold use of NLP features and propose a novel design space of features and categories incorporated in an easily usable feature canvas that is intended to help both researchers and practitioners to approach NLP endeavors in a structured manner. The resulting taxonomy simplifies the comparability of NLP features between different studies, domains, or applications and reduces the costly effort for feature engineering of practitioners and scientists for traditional ML methods. The proposed taxonomy consists of features diverse enough to demonstrate their commonalities and differences and comprehensive enough to encompass every feature I found in the systematic literature review. I believe that this diversity will encourage researchers and practitioners in IS to move beyond the scope of their discipline embracing the diversity of features that I condensed in the taxonomy. This shall increase the quality of their feature generation for research purposes or practical applications.

#### 4.1.6.1 Theoretical Contributions

From a theoretical perspective, I integrate existing literature by developing a new taxonomy, which moves beyond existing classifications and structures and groups NLP features from an integrative IS perspective. With a common classification of NLP features assuming a holistic perspective, I provide a deeper understanding of what needs to be considered when designing, evaluating, and comparing NLP applications. By combining the dimensions and characteristics of my taxonomy, I can now assist researchers and practitioners to better determine which design characteristics affect a certain outcome (e.g., for explainability and transparency post-hoc analysis). This provides opportunities for future research and more detailed insights on how to design, evaluate or compare NLP features. The taxonomy goes beyond classifying NLP-based features as it eases knowledge sharing, provides a better understanding of interrelationships among the NLP features, and thus supports decision making of researchers and practitioners (Bailey, 1994; Kundisch et al., 2021; Usman et al., 2017). The taxonomy may change the way we think and learn about NLP feature engineering in IS since it goes beyond merely classifying objects and additionally serve as a way to predict outcomes (Doty and Glick, 1994). Second, I identify and classify dimensions and characteristics beyond a disciplinary perspective of NLP feature engineering that is part of an NLP-based ML use case and its successful training and explainability analysis. Third, the various characteristics are categorized and clustered in three distinct groups (C1, C2, C3). These clusters will help researchers and practitioners to gain an overview of existing research on NLP feature engineering and to identify corresponding gaps. I believe that my research findings can offer a valuable starting point for further developing guidelines or frameworks for NLP feature engineering, especially for special use cases, in which handling accuracy issues of small data sets or conducting explainability assessments of large data models play a crucial role for successful use case deployment. Especially, this provides a profound base for technology-mediated learning to better design NLP models for ML-based learning tools, such as argumentation writing support systems or conversational argumentation learning systems. Moreover, my taxonomy may be used for educating and training novice NLP learners with a comprehensive classification of NLP features across domains and clustered groups of their applications.

#### 4.1.6.2 Practical Contributions

Given the immense growth of NLP-based ML in research and practice, as well as the interest in explainable modeling through NLP feature engineering or analysis to enhance transparency and predictive outcomes, further research on this topic is guaranteed. Such research will require a solid classification and a theoretical understanding of a design space of NLP text features. With this work, I offer such an understanding by presenting the taxonomy, its application by performing cluster analysis on it, which facilitates me

to identify white spots in literature, which are worthwhile to be investigated by future research. In this regard, I offer researchers to apply a more nuanced perspective when training NLP-based models based on my NLP feature canvas. From a practical perspective, a common understanding of NLP feature characteristics also fosters several insights for the field NLP, computer linguistics and human—computer interaction. The systematic classification of NLP features enables researchers and practitioners to design, evaluate, compare, and predict how different feature combinations impact the predictive outcome in a certain domain more effectively. For instance, at a basic level, my taxonomy determines a number of design decisions a researcher has to take when tuning an NLP model for a rare use case or to post-hoc analyze the explainability of a deep learning model. Based on my taxonomy, researchers can now identify different kinds of NLP features by combining different characteristics of my taxonomy, depending on their domain and use case.

#### 4.1.6.3 Limitations and Future Research

Of course, this research has several limitations, which provide avenues for future research. First, I only identified NLP features based on scientific publications, since my main objective was to interactively categorize studies on NLP feature engineering for structuring and enriching the IS knowledge base. Therefore, future research may adjust and extend my taxonomy based on an in-depth analysis of real-world use cases. Moreover, in the preliminary screening, I only selected research papers that explicitly mentioned text feature engineering as my goal was to achieve diversity, not volume. Future research is needed to investigate the effects of the dimensions and characteristics on performance outcomes of the NLP studies for the different application areas. Finally, based on technological advancements, more NLP features may need to be added in the future. However, encouraged by the evaluation with experts, I believe that my taxonomy represents a valuable state of the art tool for analyzing NLP features for current IS research and practice.

### 4.1.7 Conclusion

In this study, I developed a taxonomy of NLP features based on Nickerson et al. (2013)'s methodological approach and provide guidance by identifying groups and propositions based on a cluster heat map analysis. The NLP feature taxonomy is supposed to help researchers and practitioners to develop, refine, compare, and evaluate their NLP studies, e.g., when developing AI-based educational tools. I (1) illustrated the diversity of text feature engineering through a systematic literature review in which I found 133 distinct NLP features in different application domains, (2) outlined a classification framework

through a taxonomy with five dimensions and multiple characteristics, and (3) identified patterns and white spots in the application of text feature groups.

The results support the technical development for the AI-based argumentation learning systems in this dissertation. In fact, I utilize the developed feature canvas in section 5 to develop and tune NLP features for an ML-based model to identify argumentative components and relations in student-written peer reviews to provide learners with adaptive writing support. The provided foundation is, hence, necessary to concentrate the further course of this dissertation on the most important parts in human—computer interaction and IS research.

In the next section of my dissertation, I use the insights from the developed NLP features taxonomy and take a deep dive into the field of argumentation mining. Based on a systematic literature review, I built a domain agnostic model for argumentation identification, an approach which has been neglected so far at the time of conducting the study. The deep dive into argumentation identification provides a further foundation for the transformer-based architecture of the argumentation mining models developed and utilized in the course of my dissertation (e.g., in section 5).

## 4.2 Utilizing Transfer Learning for Argumentation Identification

This section further contributes to RQ1 by providing a deep dive into argumentation mining, the primary approach used in this dissertation to provide students with adaptive argumentation feedback and tutoring. The foundations of argumentation mining are outlined in subsection 2.5.

### 4.2.1 Introduction

The identification and classification of argumentation has received special attention from researchers and practitioners, since it enables the automated extraction of structured information from textual sources (Lawrence and Reed, 2019). However, at the time of conducting this study, the identification and detection of arguments was not used widely in practice, since the practical implementation of argumentation mining faces two main challenges in ML and NLP: first, even though good modeling results have been achieved in one domain, research on models that perform well for multiple domain corpora is still lacking (Lippi and Torroni, 2015b; Lawrence and Reed, 2019). Second, successful argumentation mining use cases usually depend on handcrafted NLP features that require a significant amount of manual labor and deep domain knowledge, which organizations often struggle to provide. As a result, very domain-specific models have been built that reached satisfying results in a respective domain but are mostly useless in different domain corpora and are therefore not applicable in a practical environment (Lawrence and Reed, 2019).

One promising solution avenue is utilizing transfer learning for NLP. Transfer learning is a concept in the field of machine and deep learning that tries to transfer the model learnings from one unrelated topic to another, effectively reducing the development time and increasing the prediction of the models compared to domain-specific development (Pan and Yang, 2009). Therefore, I aim to address these challenges of argumentation mining by using a novel transfer learning solution that is inter-domain applicable and does not require any labor-intensive NLP features. Current solutions in argumentation mining with unsupervised learning or classification approaches using embedding structures and neural networks (Gema et al., 2017; Xu et al., 2015) fall short of solving those issues, since they are either not generalizable or very domain specific (Lawrence and Reed, 2019).

Hence, I aim to contribute to literature and practice by presenting a novel solution that works on a deep learning model architecture and enables future scientists and researchers to build argumentation mining pipelines without intensive effort. My solution is based

on deep transformers for natural language transfer learning proposed by Devlin et al. (2018). This approach has been successfully applied in multiple NLP tasks that require language understanding with state-of-the-art results. The solution components adapted to argumentation mining that are proposed to solve the aforementioned problems are the following: First, an easy-to-train and easy-to-use transfer learning model to identify arguments and, second, a standardized corpus design to simplify new corpus introductions is suggested. To tackle the stated challenges, I develop an artifact following the CRISP-DM Model (Chapman et al. (2000), see section 3).

As I described above, I aim to develop a novel modeling approach to identify argumentative texts from multiple corpora. The CRISP-DM Model was specially built to develop data modeling approaches like ours. It consists of six different stages in which a modeling artifact is iteratively developed. At the time of conducting this research, there was no study that developed a solution to the stated argumentation mining challenges using end-to-end transfer learning approaches. In the following, I provide a brief conceptual background and present the identified challenges based on a systematic literature review following Webster and Watson (2002) and vom Brocke et al. (2015). Next, I explain the building and evaluation of the model and present and evaluate our results, followed by a discussion about the limitations and contributions of this study.

### 4.2.2 Theoretical Background

In this subsection, I am focusing on the challenges of argument identification, since this is usually the first step for an argumentation mining architecture and, thus, the foundation of every argumentation mining architecture (see subsection 2.5). Therefore, in order to analyze the current state of literature, potential research challenges, and research gaps, I conducted a systematic literature review based on Webster and Watson (2002) and vom Brocke et al. (2015). Details about the methodology of the literature review (such as search strings and data bases) are explained in the next section. I summarize the results of the very review in the following paragraph as related work on argumentation identification.

#### 4.2.2.1 Overview on Past Research on Argumentation Identification

One of the first approaches for argumentation identification was investigated by Moens (2016) which used ML techniques to develop an argument identifier for legal texts. They used heavily handcrafted features and relatively simple ML algorithms, but their results

were encouraging (Moens, 2016). Winkels et al. (2013) were among the first who experimented with unsupervised techniques in argumentation mining. Their results showed that pure unsupervised clustering does not yield satisfactory results (Sardianos et al. (2015)). Habernal and Gurevych (2015) employed the idea of word embeddings in a semi-supervised fashion for the argument identification subtask. In some cases, the results yielded a 100% improvement over previous attempts on complex online corpora. In the continuous attempt to limit the number of handcrafted features and corpus-specific knowledge, research has mainly shifted towards deep learning. A common architecture is bidirectional long-short term memory (LSTM) neural networks fed with word embeddings (Gema et al., 2017; Rajendran et al., 2018). In general, deep learning frameworks tend to give state of the art results, which approach human performance. However, the problem remains that when the model is getting introduced to an entirely new corpus, the accuracy falls significantly (Daxenberger et al., 2017). In addition, at the time of conducting this study, no effort has been made to generalize the models for new corpora, possibly because of model complexity and a high skill barrier to use (Lippi and Torroni, 2015a). As a result, there is a divide in argumentation mining between traditional ML techniques with a lot of manual labor and new deep learning techniques that require significant skill specialization. This divide is the challenge that I am addressing with the proposed model. Table 11 displays the most relevant studies and the argumentation mining tasks the respected studies contributed to. The majority of research has been using traditional ML approaches, which tend to be overly specialized. Different deep learning papers, that significantly impacted the space with their results, are also single-corpus-focused and employ a handcrafted model, which makes it hard to reproduce and use.

#### 4.2.2.2 Transfer Learning and Transformer Architecture

Unlike traditional supervised and semi-supervised ML algorithms, which assume that the distribution of the labeled and the unlabeled data is the same, in transfer learning there are no assumptions about the distributions, domain, or task (Torrey and Shavlik, 2012). Transfer learning is heavily inspired by the way humans acquire knowledge, e.g., learning how to recognize a cat can help in recognizing a tiger. In general, the study of transfer learning is motivated by the fact that people can intelligently apply previous knowledge to solve new problems (Pan and Yang, 2009). For this work, I am focusing on inductive transfer learning, where I use a model trained on a completely different corpus to do inference on another corpus for another kind of task. For this kind of transfer learning, I am using a type of recurrent neural network called “Transformer” (Devlin et al., 2018). Transformers are improving upon the LSTM-attention mechanism presented in Vaswani et al. (2017) by being able to circumvent the LSTM-attention mechanism problem of

| **Study** | **Corpus** | **Argu. Identification** | **Component Classification** | **Relation and Quality Classification** | **Modelling** |
|---|---|---|---|---|---|
| Moens et al. (2007) | Araucaria | X | | | Naïve Bayes |
| Palau and Moens (2009) | Araucaria ECHR | X | X | X | Support Vector Machine |
| Ozyurt (2012) | BioNom | X | X | | Support Vector Machine |
| Stab and Gurevych (2014a) | ComArg corpus | X | | | Support Vector Machine |
| Sardianos et al. (2015) | Greek news articles | X | X | | Conditional Random Fields |
| Zhang et al. (2016b) | Web discourse | X | | | Support Vector Machine, Naïve Bayes |
| Habernal and Gurevych (2015) | Web discourse | X | X | | Support Vector Machine |
| Poudyal et al. (2016) | Case laws (ECHR) | X | | | Support Vector Machine, Random Forest |
| Stab et al. (2018) | Web annotated text | X | X | | LSTM |

Table 11: A not exhaustive overview of approaches and level of analysis in argumentation mining literature.
Source: Own illustration

being able to process data only sequentially in the encoding step, potentially missing non-sequential information on one side of the sentence (Bahdanau et al., 2014). Transformers are able to parallelize the attention mechanism effectively, "looking" both before and after the "to-be-predicted" token. More specifically, in argumentation mining, applications of transfer learning have been lagging behind. This is in line with the observation that deep learning in argumentation mining literature has only started to appear in the past couple of years. That being said, while word embeddings as seen in Young et al. (2018) have

been used for quite some time in argumentation mining as either feature in ML algorithms or as input layers in deep learning models, no attempt has been made to use a complete transfer learning pipeline as described above.

### 4.2.3 Methodology

The first step of my research was to identify possible gaps and challenges of argumentation mining in scientific literature. Therefore, I have drawn on the approaches by Webster and Watson (2002) and vom Brocke et al. (2015). Based on well-cited literature in argumentation mining, such as Lippi and Torroni (2015a); Stab and Gurevych (2017), I identified different keywords, which researchers used to describe the pipelines of argumentation mining. Based on these, I built the following search strings trying to incorporate the previously captured namings: *("Argumentation" AND "Mining"), ("Argument" AND "Identification"), ("Argument" AND "Classification").* To find relevant literature, I applied the search string to the following six databases: AISeL, ACM Digital Library, EBSCO, IEEE Explore, ProQuest ABI Inform, and Science Direct.

The database search resulted in 12,529 hits. Titles, abstracts, and keywords are screened to fit the above-mentioned definition of argumentation mining and the application of NLP and ML to the scope of my study. I excluded papers that did not refer to a technical perspective on argumentation mining. Multiple papers are excluded due to a different research scope described in their abstract, e.g., several papers talked about argumentation in different domains, e.g., in learning science or mineral mining. Moreover, numerous studies are excluded since they are talking about classification or identification of information in completely different domains, e.g., topic classification of user-generated content or identification of user needs on Twitter.

This screening process resulted in 67 potentially relevant papers which conducted studies on argumentation mining, argumentation identification, or argumentation classification. After the elimination of all duplicates, I had 50 relevant papers left. Afterwards, a forward and backward search was carried out according to vom Brocke et al. (2015). Through screening the references, 18 studies are added to the list, resulting in 68 relevant papers, which are the basis for my related work section in the theoretical background section. I found two main literature gaps, namely, the lack of corpus-agnostic models and the problem of human-labor-intensive NLP models being costly to develop.

Next, I formulate my hypothesis that transfer learning might provide a solution to both challenges. In order to prove my hypothesis, I aim to develop a new argumentation identification pipeline based on the current state of transfer learning. Afterwards, I want

to evaluate the results of my pipeline for different corpora compared with the latest results published in the literature. In order to do so, I develop an artifact following the CRISP-DM Model (see section 3).

In the first stage, I analyzed the current state of argumentation identification achievements in literature. Second, I investigate different corpora and their results at the current state in Argument Identification in terms of precision, recall, or accuracy across multiple domains. In fact, the goal of my research is to find a model that is domain-agnostic, however, I did not find any model in the literature that achieved this goal. Third, I build an artifact that is able to identify an argument in an unknown corpus. This is achieved by using NLP are and deep learning algorithms to classify a text piece as an argument or a non-argument. The fourth stage is an iterative process of evaluation and revision of the model based on various performance metrics such as the f1-score. In this stage, I expand the model's usefulness and applicability by adding additional corpora and by identifying the best hyperparameters for my model based on the results. Finally, in the fifth stage, I draw conclusions based on the iterative process and the results.

The approach is developed using the programming language Python 3.7 for the ML applications, since it is widely known, easy to use, and supports major libraries for NLP and ML tasks. The ML-related algorithms are called from the Google-supported tool Scikit-learn. For deep learning, TensorFlow and its integrated Keras are called.

### 4.2.4 Implementation

The goal of the research is to create a model that augments the current argumentation mining techniques. Specifically, it aspires to reveal a unified model architecture that is corpus-agnostic and reduces the manual corpus-specific work. In order to accomplish this, I propose a transfer learning approach based on Deep Bidirectional Transformers for Language Understanding (BERT) as seen in Devlin et al. (2018). I hypothesize that this model architecture is uniquely suited to act as a unified, domain-agnostic, and easy-to-develop solution for argumentation mining. In order to validate the hypothesis, I conduct research structured as a CRISP-DM cycle, as it is demonstrated above. The first phase, business understanding, is explained in the introduction and the theoretical background of this section.

#### 4.2.4.1 Data Understanding

The solution to the stated problems begins with the finding of a representative corpus

for argumentation mining, followed by the construction of the model, and the training and evaluation of the classification algorithms. The first dataset acted as a blueprint for the model. The model specifications are built according to the first dataset to reduce the noise of the different annotations and assumptions of different corpora. The blueprint corpus had to be a well-structured, multi-used corpus that would minimize the possibility of corpus-specific irregularities and allow the shift of tuning and research towards the architecture and model themselves. After a thorough search of the available datasets, it became apparent that the most suitable and used corpus is the Student Essays corpus from Stab and Gurevych (2017), containing 402 annotated student essays. The corpus format has been used multiple times, the text nature (student essays) provides the best representation of a user trying to argue in a structured way. The corpus is divided into essays, each one has embedded annotated argument components "Major Claim", "Claim", "Premise". In order to deduct a binary-labeled corpus from this, the labels above are transformed into the labels "Argument" and "Non-Argument". The labels "Major Claim", "Claim", and "Premise" are labeled as "Argument", and the rest of the text is labeled as "Non-Argument". However, to achieve my goal to build a domain-crossing argumentation identification model, I needed to find additional corpora from different domains. During my systematic literature review, I decided to include the following corpora from different domains for my model:

- **AracauriaDB:** this corpus was introduced by Reed et al. (2008) and includes various sources, such as parliamentary records, news, and court summaries. The annotation quality is unknown, but it is among the very first and the most used corpora in argumentation mining.
- **User-generated web discourse:** this corpus consists of blog posts and user comments on various issues that are annotated with claims, premises, backings, and rebuttals according to the Toulmin model (Toulmin, 1984).
- **Wikipedia blog comments:** Biran and Rambow (2011) annotated two datasets. One was blog posts and their comments from LiveJournal and the second one was comments on the Wikipedia talk (WT) pages, where discussions on Wikipedia entries are made. I chose to incorporate the WT corpus since the quality is significantly better.
- **Combination:** following the standardization procedure of the proposed pipeline, all corpora have the same structure and labeling scheme. Thus, the combination of all of them was introduced to the system with the rationale of getting a holistically trained model.

| Study | Corpus | Size | Method | Metric | Result |
|---|---|---|---|---|---|
| Stab and Gurevych (2017) | Student Essays | 402 Essays | Conditional Random Fields | f1 | 85.40% |
| Moens et al. (2007) | AraucariaDB | 2800 sentences | Multinomial Naïve Bayes Classifier | accuracy | 73.75% |
| Pennington et al. (2014) | Web Discourse | 3900 sentences | Support Vector Machine | f1 | 71.40% |
| Daxenberger et al. (2017) | Wikipedia Blog Comments | 1985 documents | Convolutional Neural Network | f1 | 60.50% |

Table 12: Results of binary classification for argumentation identification based on current literature.
Source: Own illustration

#### 4.2.4.2 Data Preperation

The data preparation is split into two parts, the data standardization, and the data preprocessing. The data standardization is a meta-architecture that makes sure that each corpus abides by the same rules before entering the model. Data preprocessing is the steps that have to be taken for the corpus to be understood by the model's architecture. Since different corpora have completely different annotation schemes, in order to homogenize but also to keep tokens at the above-word length, the sub-sentence length was chosen. Each sentence was split into sub-sentences signaled by commas (',') or other punctuation ('?','!',';') and each token was assigned the "parent" label of "Argument" or "Non-Argument". All extra whitespaces, punctuation, and special characters are removed. Subsequently, since BERT requires some unique text preprocessing steps in order to work, the steps below had to be taken:

1. The labels should be converted to 0 or 1.
2. The text should be transformed to lower case.
3. The text input must be padded to be a multiple of the batch size in order to be run on the Google Cloud TPU without errors.
4. The text should be split into special sub-word tokens.
5. The text has to be converted into input features in order to be understood by the deep learning framework.

Except for converting the labels into binary, the preprocessing steps are handled by the tokenizer of BERT, minimally adapted to work on a custom classification scenario. Finally, since the classes on most corpora are very imbalanced, a balancing method was chosen. Taking into consideration that the data size is small (and the range of options that are available to avoid overfitting), the oversampling technique was used on the training sets.

#### 4.2.4.3 Modelling

After the meta-modeling and the preprocessing of the text, the result is inserted into the proposed learning architecture.

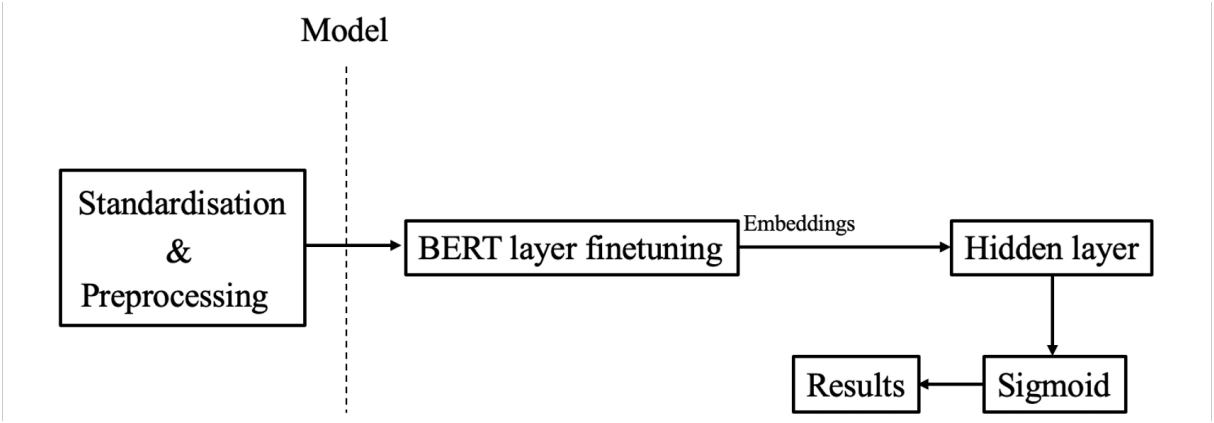



Figure 30: Outline of the proposed learning architecture for argumentation identification. Source: Adopted from Devlin et al. (2018)

Figure 30 outlines the proposed learning architecture. The goal of this model is to be a flexible, corpus-agnostic, and accurate way of identifying arguments that would be a foundation piece for the process of building an automated argument feedback system. BERT is the perfect candidate for this because of its flexibility and generalization. BERT is a pre-trained deep learning model that has been trained unsupervised using the whole Wikipedia corpus. The novelty of this architecture and what makes it ideal for transfer learning tasks is that it is able to capture semantic information from text, which can then be used for other downstream tasks without the need for retraining. BERT is able to capture this semantic information by implementing two architectural novelties. The first one breaks each word into sub-word tokens that are able to much better cope with unique words and misspellings. The second one predicts a token by looking both at the text that precedes it and the text that follows using "Transformers" instead of the traditional LSTM approach as seen in Young et al. (2018). The API access is done through Python and with the libraries of Tensorflow and Keras. The proposed architecture consists of a corpus-agnostic BERT implementation with ten fine-tuning layers with an additional single hidden layer. Fine-tuning enables to inexpensively train some of the

model's parameters, making the model specialized to argumentation mining while keeping the knowledge that the model acquired multiple days of expensive training. The last hidden layer is a recurrent neural network with 512 nodes that takes the BERT output and learns to feed into a sigmoid layer that classifies each input into one of the two classes.

Since almost all of the corpora are small, regularization techniques are needed to avoid overfitting and to stabilize the model. A dropout layer with a dropout rate of 0.7 was introduced. The whole model is considered to be quite "deep"; hence, it tends to overfit quite fast with small datasets. Thus, a high dropout rate ensures that this does not happen as fast. Also, the "callback" Keras feature was used to save the best-performing model, to decrease the learning rate once the learning has plateaued, and finally to stop the training once the evaluation metric has stopped increasing. Practically, this corresponds to the introduction of the Keras methods "ModelCheckpoint", "EarlyStopping", and "ReduceLROnPLateu". Since the goal is to determine a corpus-agnostic model, the objective of the hyperparameter tuning is to find the parameters that work best for all used corpora. A Grid Search for all corpora and all parameters would be too time-consuming, so the Student Essays corpus was used as a blueprint to significantly limit the universe of possible hyperparameters. Here, the following hyperparameters are optimized with a "sparse" grid-search: number of fine-tuned BERT layers, number of layers in the RNN, number of nodes per Layer, regularization techniques, optimizer, loss, and batch size. This search resulted in choosing the "adam" optimizer, the "binary crossentropy" loss, and the introduction of a "dropout layer" as a best practice technique. After performing a Grid Search of the student essays, the possible space was substantially reduced to the number of fine-tuned layers, since it was the main parameter with by far the most significant performance variability. Subsequently, a Grid Search of the layers of all corpora was done to reveal the optimal parameters.

#### 4.2.4.4 Evaluation

After each training, the algorithm is used to classify the posts in the previously unseen test set. The classification results (positive/negative) are compared with the true labels. The percentage of correctly classified posts compared to the total number of posts is called accuracy. Precision is the fraction of the truly positive posts among all positively classified posts, and recall is the fraction of truly positive posts that have been classified as positive. The f1-score balances between precision and recall and thus present the most suitable criterion for my use case since precision and recall are both equally important. However, I have recorded accuracy, recall, precision, and f1-score for all my classification experiments. The results are stated in Table 13.

| Corpus | accuracy | precision | recall | f1 |
|---|---|---|---|---|
| Student Essays | 80.00% | 80.41% | 91.16% | 85.19% |
| AracauriaDB | 97.46% | 95.51% | 92.37% | 93.10% |
| Web Discourse | 80.00% | 81.49% | 87.65% | 84.15% |
| Wikipedia Blog Comments | 66.59% | 70.00% | 80.75% | 74.86% |
| Combined | 75.79% | 73.22% | 80.70% | 76.57% |

Table 13: Achieved results for different domain corpora with the same BERT learning model.
Source: Own illustration

Comparing the results achieved with one unique model in Table 13 with the results for these corpora at the current state in Table 12 for argumentation identification, one can observe that my proposed model produces state of the art results (at the time of conducting the research) for almost all datasets except the student essays. The difference is minimal, and this signifies the trade-off (0.21%) between a highly specialized and complex-to-develop model versus the proposed generalized solution. It is hypothesized that AracauriaDB has the best score because it is the easiest of all; even a simple ML system (Moens et al., 2007) achieves a score of 74% accuracy. On the other end of the spectrum, it is hypothesized that the blog comments corpus has the worst score because the nature of the corpus is much less clean, and it is difficult to analyze and deduct argumentation. As a matter of fact, in Daxenberger et al. (2017) even deep learning methods fail to distinguish a claim from a non-claim in an above-chance probability.

Finally, at the time of conducting the research, there was no other published study that tried to create a generalized model out of multiple argumentation mining corpora. The combination of the corpora was made possible by creating a unified data standardization pipeline that allowed differently annotated corpora to be concatenated into a single, multi-domain corpus. The result demonstrates that even a BERT module pre-trained on non-argumentative text (Wikipedia) can produce very good results that imply language understanding at a multi-domain corpus that is structurally and semantically radically different both from the pre-trained model and from each of the sub-corpora.

### 4.2.5 Discussion

The aim of the work was to take a deep dive in to one of the challenges in argumentation mining by utilizing a novel transfer learning solution that is inter-domain applicable and does not require any labor-intensive NLP features. I conducted a systematic literature review to analyze the current state and potential gaps of argumentation mining literature. The hypothesis was that new advantages of transfer learning can solve the challenges of

domain-agnostic models and labor-intensive modeling approaches. I developed a new modeling approach utilizing BERT, and I prove that this is a suitable technology avenue for those challenges.

The contribution is twofold: first, I contribute to scientific literature by demonstrating a new modeling approach and solution to current problems of argumentation mining. The proposed model demonstrates that a whole new path has opened up for argumentation mining. Utilizing the power of transfer learning models such as BERT provides a potential solution to the fragmentation of the argumentation mining research into multiple domain-specific nodes and into a ML vs deep learning dichotomy. The research is the first step in implementing such techniques in the whole argumentation mining pipeline. Second, I contribute to practice by providing a use case on how to utilize argumentation mining in interdomain applications without intensive human modeling. Even though the proposed system uses state of the art technology to be able to infer meaning from pools of text, the availability of the toolset in an open-source and third-party maintained fashion is making it accessible to practitioners that had minimal exposure to deep learning before. In addition, the API access minimizes the time costs of running such a deep model since it is run remotely by default. The reduced time costs allow for much easier experimenting and even multiple practical applications with minimal training data input or even direct inferencing from related corpora.

### 4.2.6 Conclusion

The results of this approach guide the next chapters of my dissertation. Argumentation mining is the core approach that I utilize to provide students with adaptive feedback during their writing process. Besides using the developed NLP-feature canvas to design domain-specific features for argumentation mining in student-written texts (e.g., in section 5), I also utilize the transformer-based architecture to identify argumentation components based on my conducted research (e.g., in section 6). The next section will also build on argumentation mining and, hence, directly applies the NLP feature canvas and the findings of transfer learning for argumentation mining to provide students with adaptive writing support for peer reviews.

# 5 Designing AI-based Argumentation Writing Support Systems

This section contains one of the core contributions of my dissertation: the design and evaluation of AI-based argumentation writing support systems for students' to learn argumentation irrespective of an instructor, time, and location.

While the prior section was focusing on the technical foundations of NLP-based IS and argumentation mining approaches, this section provides a holistic overview of the requirements, the design, the development, and evaluation of AI-based argumentation writing support systems. Hence, this section contributes to the three stated research questions - RQ1b, RQ2, and RQ3a (see section 1).

**RQ1b:** *What are user and literature requirements for adaptive learning tools to support students in their argumentation learning process?*

**Method:** *Systematic literature review from an interdisciplinary perspective and user interviews.*

**Results:** *Literature issues and user requirements for the design of an argumentation learning system.*

Starting with a systematic literature review and 30 semi-structured interviews, in this section, I systematically collect theory-based and user-centered design requirements for AI-based argumentation writing support systems directly addressing RQ1b. As a result, I collect different literature issues (LI) and 180 user stories (US) representing the requirements from the knowledge-based and the field for AI-based argumentation writing support systems (see design science research in section 2).

**RQ2:** *How to design an adaptive learning tool based on Artificial Intelligence in a large-scale pedagogical scenario to support the development of students' argumentation skills in their learning process?*

Next, this section analyzes the collected design requirements and derives different design principles for the specific design of AI-based argumentation writing support systems according to Gregor et al. (2020). Addressing RQ2, the design principles are intended to guide and assist educational designers in developing new AI-based argumentation learn-

ing systems in different pedagogical scenarios to impact the development of students' argumentation skills.

**RQ3a:** *How does an adaptive argumentation writing support system need to be designed to influence students' argumentation skills in peer review scenarios?*

**Method:** *Design science research including corpus development, argumentation modeling in student-written peer reviews, mixed-method research with a series of experiments.*

**Results:** *Empirical evidence for the influence of an adaptive argumentation writing support system based on certain design principles on students' formal meta-cognitive argumentation skills and their user experience.*

Next, I aim to evaluate the design principles for AI-based argumentation writing support systems in the specific pedagogical scenario of student peer reviews. To do so, I use methods from human-centered design research, conduct annotation studies and technical experiments to model argumentation discourse structure in German student-written texts. I instantiated an AI-based argumentation writing support system based on my principles called ArgueLearn. ArgueLearn provided students adaptive feedback based on their written text by highlighting potential pitfalls in their argumentation based on a static interaction with adaptive and individual argumentation feedback. To determine the impact on students' argumentation skills, I built ArgueLearn as an instantiation of the novel system class and conducted two empirical evaluation episodes following an ex-post summative evaluation strategy (Venable et al., 2016).

In study 1, I test if individual feedback on students' argumentation helps them to write more convincing texts and thus increase their short-term argumentation skills. Hence, I rigorously compare my instantiated argumentation writing support system to a proven argumentation support tool based on an input mask (Fischer et al., 2013) in a laboratory experiment.

In study 2, I investigate if a constructive pedagogical scenario with adaptive feedback increases the long-term argumentation skills of students. Hence, ArgueLearn provides students adaptive argumentative feedback on written peer-feedbacks about business models in a large-scale lecture as a field experiment. After a period of two months and nine treatments, I then evaluate if the argumentation skills of the treatment group are higher in a second argumentation domain (societal debates based on Flender et al. (1999)) than of the control group which just received general argumentation feedback on their written peer-feedbacks.

This section's main goal is to develop a set of theoretically and empirically grounded design principles for AI-based argumentation writing support systems. This collection of design principles is an information technology (IT) artifact that establishes a general solution by depicting a class of technologies and can be instantiated into concrete IT artifacts (Iivari, 2015). Design principles serve as prescriptive statements that constitute the basis of design actions or, in other words, "how to do something to achieve a goal" (Gregor et al. (2020), p.2). Hence, I present the design principles according to their implementation, demonstrations, and evaluation of the novel class of AI-based argumentation writing support systems that provide individuals with individual argumentation support for writing tasks (e.g., a persuasive essay or a review of a peer's work). Owing to the nature of my research objective, I followed the design science methodology. In this case, the objective is to form design principles to support the systems' development of AI-based argumentation writing support systems and potentially other classes of AI-based argumentation learning systems.

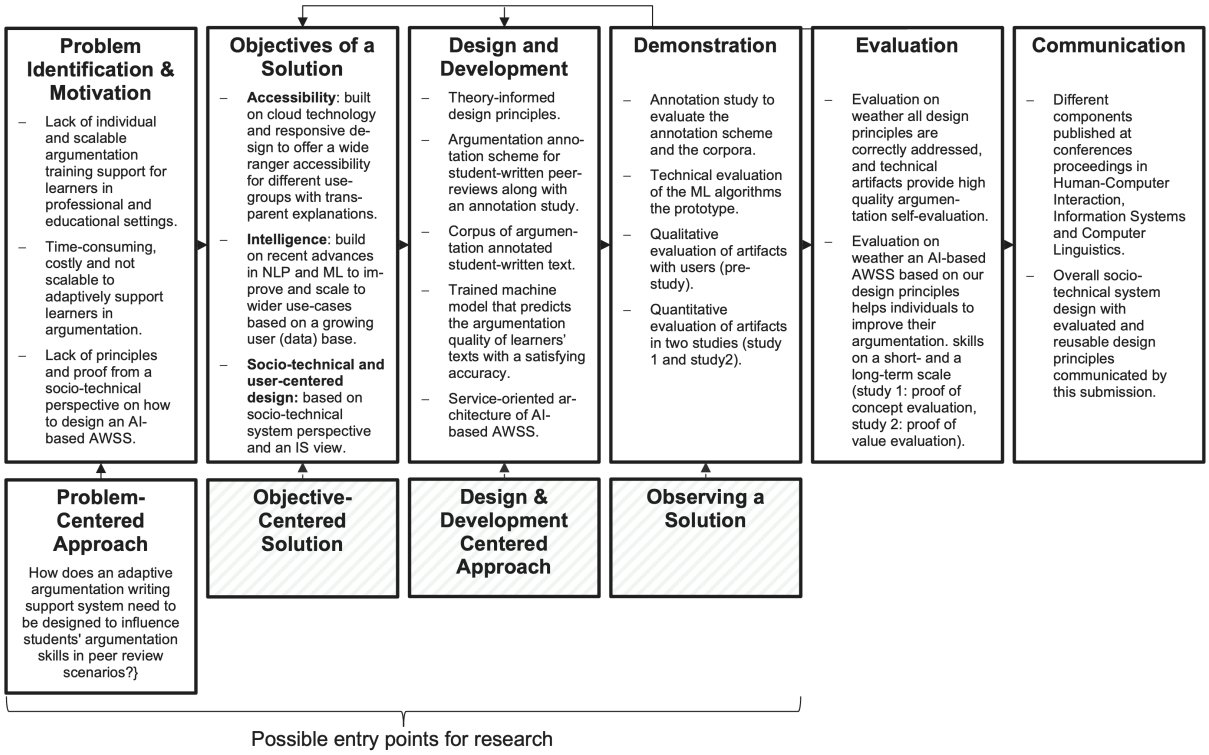



Figure 31: Design science research methodology for the design of AI-based argumentation writing support systems.
Source: Own illustration based on Peffers et al. (2007)

Figure 31 illustrates the steps taken to ensure that design knowledge is created based on insights from the application domain and the knowledge base Peffers et al. (2007).

The final prototypical system comprised a combination of artifacts based on my theory-based design principles: an argumentation annotation scheme for student-written peer

reviews along with an annotation study (artifact 1); a corpus of 1000 argumentation-annotated student-written peer reviews in German (artifact 2); a trained ML model that assesses and predicts texts' argumentation quality (artifact 3); and a fully functional AI-based argumentation writing support system (artifact 4).

Hence, this section provides a general insight into the design, development, and evaluation of AI-based argumentation writing support systems. The results (such as the design principles and the annotation scheme) influence the other succeeding sections of my dissertation (e.g., section 6).[4]

[4]The insights presented in this section are partly based on various papers (Wambsganss et al., 2020b; Wambsganss, 2021; Wambsganss et al., 2020c, 2021b, 2022b). I thank my collaborators, the reviews, and the Associate Editors for the valuable feedback on my work.

## 5.1 Requirements for the Design of AI-based Argumentation Writing Support Systems

To build a user-centered and theory-based learning system, I first derived a set of meta-requirements from the current state of scientific literature for the design of an AI-based argumentation writing support system. In this vein, I conducted a systematic literature search in the fields of human—computer interaction, IS, NLP, and educational technology using the methodological approaches of Webster and Watson (2002) and vom Brocke et al. (2015). Next, I conducted 30 semi-structured interviews with students, who are the main user group of the final artifact, using the expert interview method by Yin (2015). Based on these insights, I derived four overarching design principles based on theory and informed the design with the results form my user interviews to instantiate and evaluate the design principles in a novel AI-based argumentation writing support system. Thus, I aim not only to address the educational task (writing coherent arguments) but also to build a holistic argumentation writing support system embedded in a pedagogical scenario (structure), focusing on the learner through a user-centered and theory-motivated design (Gupta and Bostrom, 2009).

### 5.1.1 Literature Requirements for AI-based Argumentation Writing Support Systems

To derive requirements from scientific literature, I (1) defined the review scope, (2) conceptualized the topic, (3) searched the literature, and (4) analyzed the findings regarding requirements. Regarding step 1, defining the review scope, I primarily focused my research on studies that demonstrate the successful implementation of learning tools for argumentation skills. Furthermore, my goal was to identify requirements on a conceptual level with a focus on an espousal of position and representative coverage (Cooper, 1988). Regarding step 2, conceptualizing the topic, I identified two broad areas for deriving requirements: *educational technology and learning theories*. Since the creation of a learning tool for argumentation skills is a complex project that is studied by psychologists, pedagogues, and computer scientists with different methods, I first concentrated on these literature streams (Scheuer 2015). Regarding step 3, literature search, I applied different search strings such as “Argumentation Learning” or “Argumentation Mining” and applied them to the following six databases: *AISeL, ACM Digital Library, EBSCO, IEEEXplore, ProQuest ABI Inform* and *Science Direct*. Based on recent literature on technology-mediated argumentation skill learning (e.g., Stab and Gurevych (2017); Scheuer et al. (2010)), I used the following search terms: “Argumentation” AND ”Learning” OR “Argumentation” AND ”Mining” OR ”Argumentation” AND ”Identification”. Table 14 summarizes the found

| | Databases | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **AISeL** | | **ACM Library** | | **EBSCO** | | **IEEEXplore** | | **ProQuest** | | **ScienceDirect** | |
| **Search strings** | **Hits** | **Rel.** | **Hits** | **Rel.** | **Hits** | **Rel.** | **Hits** | **Rel.** | **Hits** | **Rel.** | **Hits** | **Rel.** |
| *"Argumentation" AND "Mining"* | 206 | 5 | 103 | 7 | 8 | 0 | 26 | 12 | 3446 | 3 | 206 | 5 |
| *"Argument" AND "Identification"* | 369 | 2 | 126 | 4 | 8 | 0 | 73 | 9 | 50 | 4 | 862 | 0 |
| *"Argument" AND "Classification"* | 269 | 3 | 77 | 3 | 9 | 0 | 209 | 7 | 5420 | 3 | 1062 | 0 |
| *"Argument" AND "Learning"* | 1 | 0 | 4 | 4 | 9 | 5 | 6 | 4 | 115 | 35 | 1062 | 15 |
| | With Duplicates: | | | | **130** | + | 12 Forward Search | | | | = | **138** |
| | **Without Duplicates:** | | | | **120** | | 6 Backward Search | | | | | |

Table 14: Overview on hits and relevant for AI-based argumentation learning tools in different databases.
Source: Own illustration

hits and the relevant papers for each database.

I screened the title, abstracts, and keywords of the found hits to fit the purpose of the priory stated research question and to the scope of my research. Hence, I defined specific criteria for inclusion and exclusion and reviewed the titles and abstracts of my search results in a first step. I only included literature that deals with or contributes to a kind of learning tool in the field of argumentation learning, such as an established learning theory. Several papers that deal with argumentation in other research areas were excluded. The vast majority of the found papers had no connection to both, argumentation learning and argumentation mining, and thus were excluded. In total 120 relevant papers were identified from 12,716 received hits between October and December 2018. Afterwards, a forward and backward search was carried out according to Webster and Watson (2002). By screening the references 18 of the articles were added to the list. While the result of this initial literature review impacts the entire dissertation, I summarized the most prevailing findings for the design of AI-based argumentation writing support systems in the next paragraph. Nevertheless, a general overview of argumentation learning, technology-mediated argumentation learning systems, as well as argumentation mining is provided in section 2.

I have summarized similar topics of these contributions as literature issues and formed six clusters from them. Based on these literature issues, I derived meta requirements for the design of an adaptive argumentation learning system. The literature issues as well as the relevant studies are explained and cited in the succeeding section when deriving the design principles for AI-based argumentation writing support systems to avoid repetitions.

### 5.1.2 User Requirements for AI-based Argumentation Writing Support Systems

Based on the derived literature issues and meta requirements, I conducted 30 semi-structured interviews according to Glaŝer and Laudel (2010). The interview guideline consists of 29 questions, and each interview lasted around 30 to 50 minutes. The interviewees were students at our university who are all potential users of an argumentation learning system. The participants were asked about the following topics: *experience with technology-based learning systems, perception of existing learning systems in use, the importance of skills in university education, requirements for a system that supports learning metacognition skills* (e.g., functionalities, design), *and requirements for a system that supports learning how to argue through individual writing support* (e.g., functionalities, design).

To gain impressions resulting from many years of learning experience, only master students were recruited for the interviews. The interviewed students were between 22 and 28 years old, and all students of economics, computer science, or psychology; 13 were male, 17 were female. After a more precise transcription, the interviews were evaluated using qualitative content analysis. The interviews were coded, and abstract categories were formed. The coding was performed using open coding to form a uniform coding system during evaluation (Yin, 2015). Based on these results, I gathered 180 user stories, aggregated the most common ones, and identified eight user requirements (URs) following Cohn (2004).

Students had a strong desire to practice arguing using a theoretical argumentation model in order to get learning orientation if necessary (US1). Hence, the inclusion of a theoretical learning model, e.g., the Toulmin model (Toulmin, 2003), is a major need for students (UR1). Furthermore, students expressed a desire to track present and previous performance in order to assess progress toward the preferred structure and, as a result, the learning goal (US2). All of the students agreed that the learning tool should have a straightforward, intuitive, and enjoyable user experience (UX) and be simple to set up on any device (US3), which I included in UR2. Furthermore, all students stated that they would like to have more control over the granularity of the feedback they receive on their argumentation skills (US4).

Furthermore, a large majority stated that they would like an overview of various feedback categories, such as formal and conversational structure, strengths, and shortcomings, as well as the ability to zoom in on those areas for more thorough feedback (UR3). Several times in the interviews, the content and presentation of the arguments feedback were addressed. To determine the vulnerability of the reasoning and thus precise recommen-

dations on how to improve, detailed feedback on the formal structure and linkages of their arguments is required (US5). As a result, I came up with the idea of providing feedback on formal structures and linkages of arguments in conjunction with specific improvement objectives (UR5). Following that, 80 percent of students indicated that social comparisons with peers during their argumentation learning process would obstruct their progress. Indeed, they claimed that a comparison with theoretical models would be necessary (e.g.,Toulmin (2003)) or best practices on how to improve would be more beneficial in their learning process (US7). Thus, best practices and solutions based on theory comparison should be included in the argumentation learning systems (UR6). All students agreed that the value and usefulness of the offered feedback must be particular and specific in order to continue using the tool (US8). An evaluation that is overly broad and probable recommendations that have no bearing on the individual would be a reason to avoid using the tool (UR7). These UR guided the instantiation of theory-based principles in the produced argumentation writing support system ArgueLearn, influencing the creation of design principles but mostly influencing the formulation of design principles.

After the literature has been reviewed for specific issues on AI-based argumentation writing support systems and the user requirements have been collected. The next step incorporated to derive distinct design principles for the new system class for AI-based argumentation writing support systems.

### 5.1.3 Deriving Design Principles for AI-based Argumentation Writing Support Systems

In this section, I will outline how I derive the design principles for designing AI-based argumentation writing support systems that support individuals in learning how to argue (Gregor et al., 2020). The conceptualization of the design principles was informed by theory about (argumentation) skill learning (e.g., Jonassen and Kim (2010); Pinkwart et al. (2009); Wang et al. (2020)), argumentation theories (e.g., Kuhn (1993); Toulmin (2003); Walton et al. (2008)), technology-mediated learning (e.g., Alavi and Leidner (2001); Gupta and Bostrom (2009); Hattie and Timperley (2007)), literature on argumentation mining and argumentation corpora development (Koschmann, 1996; Lippi and Torroni, 2015b; Stab and Gurevych, 2017) and the design of educational learning tools (e.g. Mayer (2014); Scheiter and Gerjets (2007)).

According to design science research, I based the formulation of my design principles on a particular kernel theory to guide the development of AI-based argumentation writing support systems. As introduced in section 2, I relied on self-regulated learning theory (Bandura, 1991; Zimmerman and Schunk, 2001) as an overarching kernel theory and,

in particular, on self-monitoring and self-evaluation based on formative feedback (Bandura, 1991; Hattie and Timperley, 2007). However, I do not aim to focus solely on self-regulated learning and the learner's entire cognitive process but rather on the proven self-reliant learning so that individuals can learn argumentation skills irrespective of instructor, time, and location. Being aware that self-regulated learning involves more than formative feedback, self-monitoring, and self-evaluation, I am still focusing on these (individual) concepts—in combination with other guiding literature—with the simple objective of designing a novel argumentation learning system to better support humans in learning argumentation skills based on recent technological advances embedded in known theoretical concepts (such as self-regulated learning). Below, I will further outline how I use self-regulated learning in combination with other theories to derive four design principles to achieve my research objective.

**Transparent and In-Depth Skill Feedback Based on Learners' Individual Errors**

Self-regulated learning theory holds that humans learn better through formative feedback and goal setting (Bandura, 1991; Zimmerman and Schunk, 2001). Critical reflection through self-monitoring and self-evaluation is particularly important for students' effective learning (Zimmerman and Schunk, 2001). This may be an initial trigger for the individual's learning process and, thus, the creation of new knowledge structures. However, the correct proportion of self-evaluation in combination with a learning goal is important for humans to learn effectively (Bandura, 1991; Zimmerman and Schunk, 2001). Therefore, I propose that the right level of feedback on learners' skills, such as argumentation, could lead to greater self-efficacy and increased motivation to learn how to argue in a structured and well-formed way (Bandura, 1977). According to Bjork et al. (2013), accurate self-monitoring and evaluation of one's learning progress are fundamental components for effective self-regulation of learning. Nevertheless, humans struggle to monitor and evaluate their learning and comprehension of complex materials (Bjork et al., 2013). Hence, the design of appropriate monitoring and evaluation elements for certain skills might help learners to continously learn more effectively from a long-term perspective (i.e., Roediger and Karpicke (2006)). One way to facilitate effective monitoring and evolution of learning is to combine skill monitoring with formative assessment and performance feedback (Black and Wiliam, 2009; Hattie and Timperley, 2007). From a learning theoretical perspective, individual formative feedback is crucial for the learning of metacognition skills, such as argumentation (Vygotsky, 1980). To be effective, feedback should define goals, monitor progress toward the goals, and identify the activities that will bring the learner toward the goals (Hattie and Timperley, 2007). Furthermore, research has suggested that feedback on learners' errors can greatly facilitate new learning (Metcalfe, 2017). According to Metcalfe (2017), p. 484, making errors and receiving feedback "enhances later memory for and generation of the correct response, facilitated active learning, [and] stimulates the

learning to direct attention appropriately."

Adaptive feedback has been suggested as a means of monitoring students' argumentation skills (Afrin et al., 2021; Noroozi et al., 2018; Scheuer et al., 2010; Lawrence and Reed, 2019). Accurate and detailed argumentation writing feedback is most helpful and effective, as evidenced by improvements in writing outcomes (Afrin et al., 2021). Research has also found that transparent argumentation highlighting (e.g., with in-text colors) in combination with background information and explanations is more effective in helping users understand the feedback (Afrin et al., 2021; Wang et al., 2020). In particular, colored in-text feedback on claims and premises has been shown to successfully support argumentation skills (Afrin et al., 2021; Chernodub et al., 2019; Zhang et al., 2016a). However, adaptive recommendations and explanations have been repeatedly mentioned as a requirement for users to successfully understand and incorporate feedback (Afrin et al., 2021; Hattie and Timperley, 2007; Xu et al., 2021).

Hence, based on Gregor et al. (2020), I derive the following design principle (DP1):[5]

**DP1:** *Employ transparent and in-depth skill feedback based on learners' individual errors to facilitate self-evaluation during their individual learning process.*

### Learners' Choice of Different Granularity Levels, Visualizations, and Representation of Feedback

Providing learners with greater control over their learning process can lead to positive impacts on learning effectiveness and outcomes (Brown et al., 2016). According to Howardson et al. (2017), the concept of learner control involves understanding both learner- and instruction-centric constructs. Hence, it is important to incorporate not only surface-level learner control (e.g., when and where the exercise is completed) but also deep-level learner control (e.g., learners' capacity to control the level of feedback information) (Brown et al., 2016). In their active learning approach, Bell and Kozlowski (2008) identified the concepts of guided exploration and error management as most effective for self-regulated learning. Guided exploration allows learners to act on and receive feedback from a digital learning tool, whereas error management motivates educational designers to develop pedagogical scenarios in which learners can make errors and inductively discover strategies to improve and learn accordingly (Brown et al., 2016; Zacher and Frese, 2018). Both are typically combined with personal suggestions and explanations.

[5]Please note that all my design principles aim to support IS designers in designing AI-based argumentation writing support systems for individuals to improve their argumentation skills autonomously and independently of an instructor and irrespective of time, and place.

A high degree of surface-level and deep-level control of learning feedback has been shown to positively impact skill-based learning outcomes (Carolan et al., 2014; Kraiger and Jerden, 2008). Based on cognitive load theory (Sweller, 1994), Mayer (2003) also identified the learners' control principle as a major element in their e-learning theory. It aims to enable learners to adjust the amount of input information needed for their personal learning process. Moreover, in his cognitive theory of multimedia learning, Mayer (2009) identified "multimedia principles," including that "people learn more deeply from words and pictures than from words alone" (Mayer (2009), p.47). He assumed that a learner has two separate information processing channels (auditory—related to reading texts, and visual—related to comprehending visualizations). Each channel has limited capacity, and learning is an active process of filtering information based on prior knowledge. Therefore, to better guide learners, an argumentation learning system must incorporate both words and images to reduce the load on a single processing channel. Providing representations of argumentation structures to humans is also reflected in the theory of representational guidance for argumentation skill learning (Scheuer et al., 2010; Suthers, 2003).

Accordingly, argumentation representations (e.g., a visualization of an argumentation discourse as linked graph structures) foster argumentation by stimulating and improving individual reasoning and learning (Suthers, 2003; Suthers and Hundhausen, 2001).

For AI-based argumentation writing support systems, learner control plays a crucial role in feedback allocation design. I thus derive the following second design principle:

**DP2:** *Offer a selection of different granularity levels, visualizations, and representations of the feedback to provide the learner with the right amount of feedback information necessary to foster self-controlled learning.*

### Machine Learning Model on a Validated and Sufficient Corpus Retrieved from a Related Text Domain of the Pedagogical Scenario

The main objective is to provide learners with adaptive and individual feedback that enables them to monitor and evaluate their current argumentation skill levels irrespective of instructor, time, and location. While argumentation feedback can also be achieved through rule-based approaches (Huang et al., 2016), by ontologies or syntactical structures (Scheuer et al., 2012b), and human-based correction (Afrin et al., 2021), novel argumentation mining algorithms offer scalable ways to provide adaptive feedback based on the sentence, or even the sub-sentence level on unfamiliar text domains (Chernodub et al., 2019; Visser et al., 2021; Zhang et al., 2016a). However, the availability of high-quality labeled corpora is crucial in designing, training, and testing argumentation mining algorithms. Numerous studies have dealt with the creation of annotated datasets but these

have all been limited to particular genres, ranging from well-structured legal (Palau and Moens, 2009) and scientific documents (Kirschner et al., 2015; Lauscher et al., 2019), to ambiguous and less formal social web content Habernal and Gurevych (2015). The creation of gold standards and test collections requires a formal representation model and corresponding annotation guidelines. While several well-defined models exist in argumentation mining (Freeman, 2001; Toulmin, 2003; Walton et al., 2008), no general argumentation annotation scheme is applied across all textual domains and genres. The proposed representations differ in granularity, expression power, and categorization (Lawrence and Reed, 2019; Visser et al., 2021). Therefore, studies with several annotators when introducing new annotation schemes are crucial for the quality of argumentation corpora (Visser et al., 2021). Hence, it is important to ground new argumentation schemes not only on argumentation mining literature but also on theoretical concepts, such as defined argumentation models (Toulmin, 2003). In-text feedback on the sentence level is known to be more effective than higher-level argumentation feedback (e.g., on the paragraph or document level) (Afrin et al., 2021). Hence, for AI-based argumentation writing support systems, a corpus based on a theory-grounded annotation scheme is necessary to reliably capture argumentation discourse at the sentence level (Stab and Gurevych, 2014a, 2017; Visser et al., 2021). In general, when collecting a dataset to train a model for real-life user feedback, it is necessary to ensure that the data are based on a representative and balanced subset of the domain in which the model is later employed (Pustejovsky and Stubbs, 2013).

To develop an accurate AI-based argumentation writing support system, it is necessary to ensure that the annotation scheme, as well as the collected corpus, are related to—if not retrieved and designed for—the specific pedagogical scenario in which the learning tool is deployed. Moreover, the corpus should be of a sufficient size to use trained models in real-world scenarios. For sufficient corpus quality, a validation study for the particular annotation scheme and the domain is also necessary to control for the objectivity of the feedback the students later receive (Visser et al., 2021). This is achieved by following rigorously defined annotation guidelines for guiding the annotators toward an adequate agreement (Stab and Gurevych, 2017; Visser et al., 2021). Finally, the corpus should then be used to train, validate, and tune a machine learning model to provide accurate and useful feedback (Bird et al., 2009). Here, it is important to ensure that a reliable prediction of the annotated argumentation structures is possible based on accuracy, recall, and precision (Bird et al., 2009).

All in all, based on the rather technical requirement for creating an argumentation annotation scheme, a corpus of sufficient size and quality, and a tuned ML, I derive the following design principle for the design of AI-based argumentation writing support systems:

**DP3:** *Train a machine learning model on a balanced and representable corpus retrieved from a related text domain of the pedagogical scenario. Ensure that the corpus is of sufficient size and annotated based on rigorous annotation guidelines from argumentation theory and validated for objectiveness by multiple human annotators to ensure that humans receive individual and useful argumentation feedback.*

**Transparent User Interaction Incorporated throughout the Entire Argumentation Learning Process**

Based on the user requirements, I noted that learners aimed to receive transparent explanations about the argumentative writing goal, theory embedding, and a clear learning goal that guides them before, during, and after the argumentative writing task.

The user requirements can be explained from a socio-technical systems perspective. Hence, I derived that for a holistic and interdisciplinary artifact design for an AI-based technology-mediated IS for skill learning, according to Gupta and Bostrom (2009), it is relevant to follow a socio-technical systems perspective (Bostrom and Heinen, 1977). For the tool's long-term adaption and success, it is important not only to address the educational task (writing correct arguments) with a technology-mediated learning tool but also to build a holistic argumentation learning system embedded in a pedagogical scenario (structure) and focus on the learner through a user-centered and theory-motivated design. To provide general guidance and facilitation in the long-term learning process, and for argumentation learning, in particular, the concept of technology-enhanced scaffolding offers theoretical backing (Gupta and Bostrom, 2009; Sharma and Hannafin, 2007).

Scaffolding was defined by Wood et al. (1976) as temporary instructional help for learners to overcome barriers within their zone of proximal development, modifying learners' specific learning paths and experiences. Scaffolding, which has its roots in the social constructivist theory (Vygotsky, 1980; Wood et al., 1976), asserts that intersubjectivity between the instructional designer and the individual learner, as well as between learners, is essential for learning. Scaffolds support learning processes by giving advice on how to use relevant methods and structures (Gupta and Bostrom, 2009), structure tasks (Reiser, 2004), helping to self-monitor learning processes (Lehmann et al., 2014), providing cues or hints on how to complete tasks, and marking critical features related to the task while also highlighting discrepancies (Reiser, 2004). In his theory of learner-centered design, (Soloway et al., 1994) named the concept of scaffolds a central component of learning software when the task is to complete constructive activities, such as writing argumentative texts. Accordingly, an argumentation learning system should comprise a specific goal, purpose, and orientation to help learners reflect on what they are learning and give

guidance on the context and their tasks (Soloway et al., 1994). Hence, I conclude with the following fourth design principle, based on Gregor et al. (2020):

**DP4:** *Employ a mechanism that explains content and goals before, during, and after the learning activity to monitor the user's learning process actively and transparently and to convey a goal and purpose of learning to ensure long-term learning success.*

After deriving the design principles, the next section proposes the design and implementation of an architecture for the system class of AI-based argumentation writing support systems. The design and implementation encapsulate the four derived design principles and guide the demonstration of an operational system.

### 5.1.4 Conclusion

In this subsection, I described how I derived different requirements from the knowledge base (scientific literature) and the field (users) for the design of AI-based argumentation writing support systems. The screened literature issues, as well as the user stories, informed the conceptualization of four novel and distinct design principles for the design of AI-based argumentation writing support systems. The design principles are addressed in the next section by proposing an service-oriented architecture for this class of argumentation learning systems.

Therefore, with this subsection, I have contributed to the lack of design knowledge on argumentation learning systems. Moreover, I contributed to answering RQ1b and RQ2 stated in section 1. In the next section, I will outline how I annotated and developed a novel corpus, which serves as an underlying necessity for the proposed architecture and the demonstration as well as the evaluation of AI-based argumentation writing support systems. After explaining the corpus collection, I will outline how I answered RQ3a by instantiating an AI-based argumentation writing support system called ArgueLearn and by evaluating the design principles in two consecutive studies.

## 5.2 Design and Implementation of an AI-based Argumentation Writing Support System

This section encompasses the design and implementation of an AI-based argumentation writing support system called ArgueLearn. Based on the four design principles, I conceptualized the architecture of an AI-based argumentation writing support system based on a service-oriented architecture approach (Niknejad et al., 2020). The architecture consists of four artifacts that form the novel class of AI-based argumentation writing support systems. All artifacts together instantiated the theory-derived design principles. The specific instantiation was informed by the 30 semi-structured interviews with students to form a concise approach and system class of AI-based argumentation writing support systems (Hevner et al., 2004). The interviews eventually enriched the design principles with knowledge from the field and the end users and helped us build a student-centered learning tool. The final argumentation writing support systems ArgueLearn allows users to enter a text and receive individualized and instant feedback on their argumentation. A screenshot of ArgueLearn with examples of functions (e.g., F1–F6) is presented in Figure 39. Hence, in this section, I will firstly outline the proposed service-oriented architecture of AI-based argumentation writing support systems. Afterwards, I will explain the corpus annotation and development process. Finally, I will introduce the reader to the human–computer interaction of the final version of ArgueLearn.

### 5.2.1 Proposed Architecture of AI-based Argumentation Writing Support Systems

An service-oriented architecture system is composed of multiple modules with a set of defined functionalities. In this dissertation, I refer to the modules as artifacts, as described in design science research (Gregor and Hevner, 2013). The artifacts can operate and be updated independently. Thus, an service-oriented architecture allows for high availability and interoperability for different domains, use-cases, and user groups (Niknejad et al., 2020). This fosters the reproducibility and transparency of the artifacts for adoption to other use cases and domains (e.g., for other languages or subjects). Figure 32 provides an overview of the proposed architecture.

The architecture of an AI-based argumentation writing support system consists of four artifacts that form each novelties in their distinct research areas and together form a new system class. The core of the architecture is the pedagogical scenario, in which learners conduct a persuasive writing exercise and receive ongoing and individual argumentation self-evaluation through an AI-based argumentation writing support system irrespective

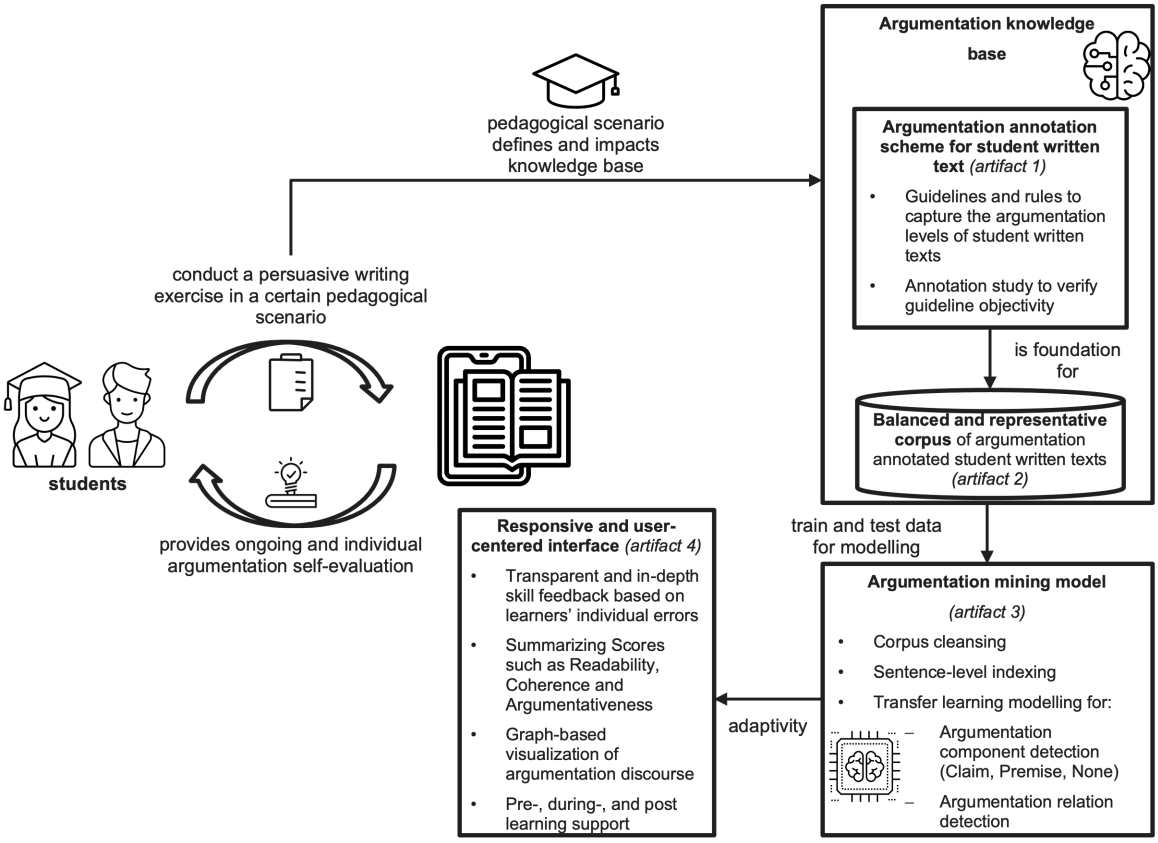



Figure 32: Overview of the proposed architecture of an AI-based argumentation writing support system.
Source: Own illustration

of instructor, time, and location. The pedagogical scenario defines and impacts the argumentation knowledge base. The knowledge base comprises an argumentation annotation scheme for students' written text with guidelines and rules to capture the students' argumentation levels in the particular pedagogical scenario (artifact 1). The evaluated annotation scheme is the foundation of an argumentation annotated corpus of train and test data (artifact 2) for the argumentation mining algorithm. The argumentation mining module (artifact 3) is trained on the corpus utilizing a machine learning approach to access the individual argumentation level in student-written texts. The novelty of machine learning compared to argumentation skill feedback in the literature (i.e., Huang et al. (2016); Scheuer (2015); Stab and Gurevych (2017)) arises from its broader applicability, scalability, and capacity to improve based on ongoing usage through the generation of new training data and self-learning capabilities (Lawrence and Reed, 2019). Finally, the responsive and student-centered interface enriches the students' self-evaluation with individualized, transparent, and in-depth skill feedback (artifact 4).

All artifacts together instantiated the four derived design principles. The specific instantiation was informed by the 30 semi-structured interviews with students to form a concise approach and system class of AI-based argumentation writing support systems

(UR described at the beginning of the prior section).

In the following, I will outline the annotation and evaluation of a novel corpora of 1000 student-written peer reviews, which served as the adaptive feedback algorithm for the instantiation of the proposed architecture of ArgueLearn.

### 5.2.2 Annotating and Developing a Corpus of Argumentation Annotated Student-Written Peer Reviews

In this section, I will outline how I derived a novel annotation approach to capture claims and premises of arguments and their relations in student-written persuasive peer reviews on business models in German language. I propose an annotation scheme based on annotation guidelines that allow modeling claims and premises as well as support and attack relations for capturing the structure of argumentative discourse in student-written peer reviews. Moreover, I conduct an annotation study with three annotators on 50 persuasive essays to evaluate the annotation scheme. The obtained inter-rater agreement of $\alpha = 0.57$ for argument components and $\alpha = 0.49$ for argumentative relations indicates that the proposed annotation scheme successfully guides annotators to moderate agreement.[6]

#### 5.2.2.1 Introduction

To design, train, and test argumentation mining algorithms, the availability of high-quality labeled corpora is crucial. Therefore, numerous prior works have dealt with creating annotated data sets. However, they are all limited to a particular genre, ranging from well-structured legal (Palau and Moens, 2009; Ashley and Walker, 2013) and scientific documents (Kirschner et al., 2015; Houngbo and Mercer, 2014) to rather ambiguous, vague, and less formal social web content (Wachsmuth et al., 2014; Aharoni et al., 2014; Cabrio and Villata, 2014; Habernal and Gurevych, 2015). Corpora that are applicable to the design and development of argumentative writing support systems are scarce (Stab and Gurevych, 2017). At the time of conducting this research, the only collection from the education domain that is annotated for argumentative discourse structures was presented in Stab and Gurevych (2014a). It is composed of 90 persuasive essays written by students in English and later extended to include 402 essays (Stab and Gurevych, 2017).

However, these corpora are (1) annotated in English only and (2) not derived from a

[6]The insights presented in this section are partly based on the paper Wambsganss et al. (2020c). I thank my collaborators, the reviews, and the Associate Editors for the valuable feedback on my work.

specific learning scenario that would leverage the effective use for an argumentative writing support system. Consequently, there is a lack of linguistic corpora for training models that provide students with adaptive feedback about the quality of their argumentation in common scenarios in large-scale lectures or the growing field of massive open online courses (Seaman et al., 2018), such as peer reviews where students give each other argumentative feedback on a specific task, e.g., on a previously developed business model (Rietsche and Söllner, 2019).

Creating gold standards and test collections requires a formal representation model as well as corresponding annotation guidelines. In this work, I introduce an argumentation annotation scheme for persuasive student-generated peer reviews extracted from a common learning scenario. Moreover, I present a corpus of 1,000 student-written peer reviews that are annotated for argumentation components and their relations. The contribution is threefold: (1) I derive an annotation scheme for a new data domain for argumentation mining based on argumentation theory and previous work on annotation schemes for persuasive student essays (Stab and Gurevych, 2017, 2014a), (2) I present an annotation study based on 50 persuasive peer reviews and three annotators to show that the annotation of student peer reviews is reliably possible, (3) I present a final and freely available corpus of 1,000 student peer reviews collected in a lecture about business innovation in German.

This corpus as well as the underlying annotation scheme lays the foundation for the design and development of AI-based argumentation writing support systems and conversational argumentation learning systems in the course of my dissertation.[7]

#### 5.2.2.2 Theoretical Background

##### Argument Annotated Corpora and Annotation Schemes

As Lawrence and Reed (2019) state, one of the challenges faced by current approaches to argument mining is the lack of large quantities of appropriately annotated arguments to serve as training and test data. Since the availability of labeled corpora is crucial for designing, training, and evaluating argumentation mining algorithms, numerous prior works have dealt with creating annotated data sets, such as the Araucaria corpus (Reed et al., 2008), the European Court of Human Rights (ECHR) corpus of Mochales and Moens (2008), the Debatepedia corpus (Cabrio and Villata, 2012), or the persuasive essays corpus of Stab and Gurevych (2014a) with 90 essays and the corpus of Stab and Gurevych (2017) with 402 persuasive student essays. These corpora have been widely used

[7]The annotation guidelines can be found in the appendix of this dissertation.

for various argumentation mining tasks, such as the identification of argument components (Rooney et al., 2012), corpus wide argumentation mining (Ein-Dor et al., 2019) or end-to-end argumentation mining (Persing and Ng, 2016).

Creating gold standards and test collections requires a formal representation model as well as corresponding annotation guidelines. While a number of well-defined models exist in the field of argumentation mining (Freeman, 2001; Perelman, 1971; Pollock, 1995; Walton and Macagno, 2015; Walton, 1996; Wambsganss et al., 2020a), there is no general argumentation annotation scheme across all domains and genres of texts. Instead, the proposed representations differ in granularity, expression power and categorization (Lawrence and Reed, 2019). Therefore, conducting annotation studies with several annotators when introducing new annotation schemes is crucial for the quality of argumentation corpora.

**Argument Annotated Corpora in Education**

With the exception of the corpora proposed in Stab and Gurevych (2014a) and Stab and Gurevych (2017), previously presented argument annotated datasets are not easily applicable for the development of argumentative writing support systems for students in a real-world case, since they are (1) not extracted from an educational learning scenario in which the annotation could be used for training a model that provides students feedback on the texts, and (2) often not annotated at the level of discourse (Stab and Gurevych, 2017; Lawrence and Reed, 2019), which is necessary for example to give students feedback on insufficiently supported claims. Stab and Gurevych (2014a) identified the lack of linguistic corpora in the domain of student-written texts for designing and developing argumentative writing support systems for adaptive feedback by leveraging argumentation mining techniques (Stab and Gurevych, 2014a). Therefore, they introduced an annotation scheme for annotating argument components and their relationships in persuasive student essays. Afterwards, several studies built on their corpus, including e.g., Carlile et al. (2018) who take a subset of the essays and annotate their persuasiveness, or Ke et al. (2018) who train a persuasiveness scoring model on them. However, the transfer of argumentation corpora to other educational domains (e.g., common learning scenarios such as peer reviews) and other languages falls short in current literature.

**Essay Scoring Corpora**

Besides annotated datasets of argumentative student-written texts, several corpora exist in the field of automatic essay scoring. The goal of this task is to automatically rate textual documents in the form of holistic scores based on their content and form (Horbach et al., 2018). Most corpora are built on student-written content, e.g. the Cambridge Learner Corpus (CLC) (Yannakoudakis et al., 2011) with 1,244 English essays, the Swedish high school corpus with 1,702 essays (Östling et al., 2013), or the TOEFL11 corpus with 1,100

English essays written by language students (Blanchard et al., 2013). However, the corpora are usually annotated with a holistic score on the level of full documents only, e.g., *low, medium, high* in the TOEFL11 corpus, while specific argumentation structures are commonly ignored. In fact, the persuasiveness of essays is — if at all considered — usually only regarded as one sub variable of the overall document (e.g., in the annotations of Östling et al. (2013) and Yannakoudakis et al. (2011)). Consequently, these corpora lack applicability for the development of argumentative writing support systems. The argumentation quality is often only annotated as a qualitative score on a 1 to 3 range (e.g., Horbach et al. (2018)) and therefore not sufficient enough to train a sophisticated supervised machine learning model for argumentative writing assistants. Corpora from the field of automatic essay scoring usually neither distinguish between different types of argument components (e.g., claims or premises) nor are they built on a rich argumentation annotation scheme. Nguyen and Litman (2018) demonstrated the value of argumentation mining for automated persuasive essay scoring by evaluating different argumentation mining features for improving essay scores. However, essay scoring corpora mostly do not focus on the annotation of argumentation relations, and therefore, disqualify for a foundation for sophisticated models for feedback on argumentative discourse through writing support systems (e.g., feedback on unsupported claims). I aim to address this literature gap by presenting and evaluating an annotation scheme as well as an annotated argumentation corpus built on student-written texts with the objective of enabling researchers to develop novel argumentative writing support systems for students.

#### 5.2.2.3 Corpus Construction

I compiled a corpus of 1,000 student-generated peer reviews in which students provide each other feedback on previously developed business models. Peer reviews are a modern learning scenario in large-scale lectures, enabling students to reflect on their content, receive individual feedback from peers, and thus deepen their understanding of the content (Rietsche and Söllner, 2019). Moreover, they are easy to set up in traditional large-scale learning scenarios or in the growing field of distance-learning scenarios such as massive open online courses (Seaman et al., 2018). This can be applied to train skills such as argumentation. However, since not many suitable corpora are available to provide argumentation feedback that (a) contain annotated persuasive student peer reviews, (b) consist of a sufficient size to be able to use trained models in a real-world scenario, and (c) follow an annotation guideline for guiding the annotators towards an adequate agreement, I propose a new annotation scheme to model argument components as well as argumentation relations that reflects the argumentative discourse structure in persuasive peer reviews. I based the annotation scheme on the model of Toulmin (1984) and the

studies of Stab and Gurevych (2014a, 2017).

To build a reliable corpus, I followed a four-step methodology, illustrated in Figure 33: (1) I examined scientific literature and theory on how to model argumentation discourse structures in texts from different domains, (2) I randomly sampled 50 student-generated peer reviews and, on the basis of the findings from literature and theory, developed a set of annotation guidelines consisting of rules and limitations on how to annotate argumentation discourse structures, (3) I applied, evaluated and improved the guidelines with three native-speakers in three consecutive workshops to resolve annotation ambiguities, (4) I applied the final annotation scheme based on the 15 pages guidelines to a corpus of 1,000 student-generated peer reviews.[8]

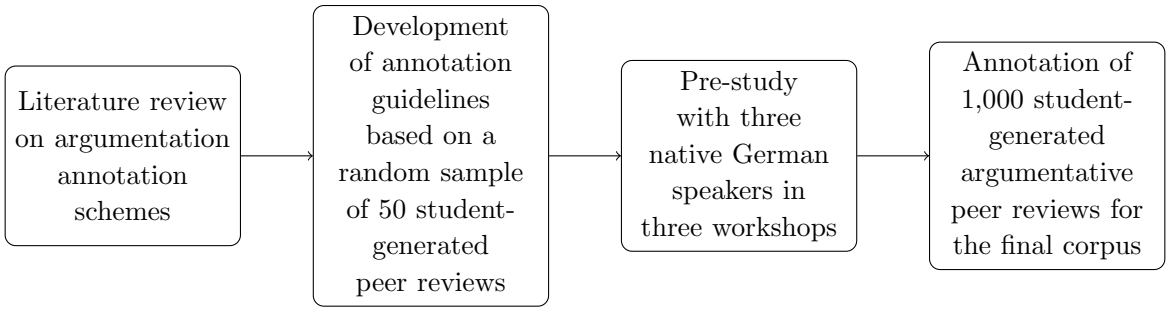



Figure 33: Overview on the process of collecting a corpus of 1,000 annotated peer reviews written by students.
Source: Own illustration

### Data Source

I gathered a corpus of 1,000 student-generated peer reviews written in German. The data was collected in one of the mandatory business innovation lectures in a master program at the university. In this lecture, around 200 students develop and present a new business model for which they receive three peer reviews each. Here, a fellow student from the same course elaborates on the strengths and weaknesses of the business model and gives persuasive recommendations on what could be improved. I sampled a random subset of 1,000 of these reviews out of around 7,000 documents collected between 2014 and 2018.

### Annotation Scheme

The objective is to model the argumentation discourse structures of student-generated peer reviews by annotating argumentation components and their relations. Most of the peer reviews in the corpus follow a similar structure. They describe several strengths or weaknesses of the business model under consideration, backing them up by examples, statistics, intuitions, or citations. These strengths and weaknesses can also be formulated

[8]The annotation guidelines, as well as the entire corpus, can be accessed at https://github.com/thiemowa/argumentative_student_peer_reviews.

to make a certain recommendation for improvements to the business model. An argumentation component, e.g., a strength, weakness, or suggestion is only regarded as a "*claim*" if it contains a certain standpoint, which can also represent a complete sentence. The basic annotation scheme is illustrated in Figure 34.

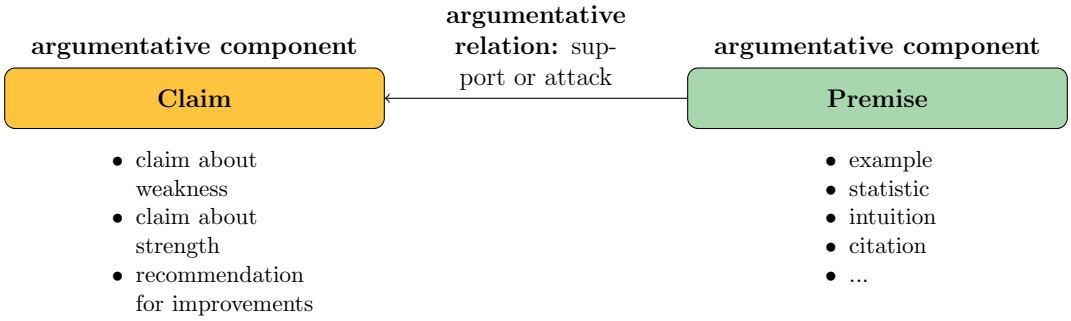



Figure 34: Argumentation annotation scheme including argument components (claim, premise) and argumentative relations (support or attack).
Source: Own illustration

### Argument Components

For argumentation components, I follow established models in argumentation theory which provide detailed definitions of argument components (e.g., Toulmin (1984); Walton et al. (2008); Weinberger and Fischer (2006); Perelman (1971); Pollock (1995); Walton (1996); Freeman (2001); Walton and Macagno (2015); Van Eemeren and Grootendorst (2016)). These theories generally agree that a basic argument consists of multiple components and that it includes a *claim* that is supported or attacked by at least one *premise*. Also in the domain of student-written peer reviews, I found that a *claim* is the central component of an argument. It is a controversial statement (e.g., claiming a strength or weakness of a business model - see examples below) that is either true or false and should not be accepted by the receiver of the feedback without additional support or backing. The *premise* supports the validity of the *claim* (e.g., by providing a statistic, quote or a value-based intuition). It is a reason given by the author to persuade the reader of her *claim*. Below are two examples of *claims* from the corpus (*1.* being a strength, *2.* being a weakness) and their supporting *premises*.[9]

1. "The value proposition is very well done.claim (strength) *It is short and concise and the advantages or benefits are well highlighted.*premise (example) *As a customer I would like to try the product after reading it.*"premise (intuition)

2. "Unfortunately, the value proposition canvas is very poorly filled in.claim (weakness)

[9]Since the original texts are written in German, I translated the examples to English for the sake of this work.

*The points are far too little elaborated and far too little described or explained. The customer jobs should be described much more precisely.*" premise (example)

### Argumentative Relations

The basic argumentation structure in the corpus of student-generated peer reviews consists of several claims, each independently supported by one or more premises. Since in the data domain weaknesses and strengths of a business model are discussed, the texts do generally not present a major claim as is the case for example in English essays annotated by Stab and Gurevych (2014a) and Stab and Gurevych (2017). Instead, the documents I deal with consist of a set of claims supported by one or more premises. However, premises not only *support* a statement but may also *attack* a claim, e.g., when used as a stylistic device or to illustrate uncertainty in the argumentation. Hence, more complicated constellations of claims and premises are possible, in which a claim is supported by several different premises or by a chain of premises, in which each premise is in turn supported by another premise. In the same way, a claim can be supported by one premise and attacked by another or supported by a premise that is attacked by another premise. Nevertheless, the simplest form consists of a claim supported by a single premise. To provide an overview, I illustrated three basic examples of annotated relations in the corpus (see Figure 35). Some statistics on the occurrences of different patterns in the dataset can be found in Table 20.

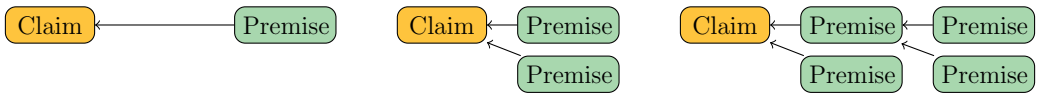



Figure 35: Examples of argumentation relations that could be found in the corpus. The arrow represents a relationship, which might be support or attack. The rectangle signifies claim or premise, which is an argument component.
Source: Own illustration

### Annotation Process

Three native German speakers annotated the peer reviews independently from each other for *claims* and *premises* as well as their *argumentative relationships* in terms of support and attack, according to the annotation guidelines I specified. Inspired by Stab and Gurevych (2017), the guidelines consisted of 15 pages, including definitions and rules for what is an argument, which annotation scheme is to be used and how argument components and argumentative relations are to be judged. Several private training sessions and three team workshops were performed to resolve disagreements among the annotators and to reach a common understanding of the annotation guidelines. We used the brat rapid annotation tool, since it provides a graphical interface for marking up text units and linking their relations (Stenetorp et al., 2012). After the first 50 reviews were annotated by

all three annotators, we calculated the inter-annotator agreement scores. As we obtained satisfying results, we proceeded with a single annotator who marked up the remaining 950 documents. Following Stab and Gurevych (2014a), we annotated argument components in terms of claims and premises. The specific component types, e.g., *weakness, strength* for claims or *intuition, statistic* for premise were not annotated in this study. However, I believe this would be a useful addition to future work. In Figure 36, I display an example of an entire peer review with the corresponding annotations.

#### 5.2.2.4 Corpus Analysis

I analyzed the results of the annotation process in order to examine (1) the *reliability of the corpus* and (2) the *major disagreements* in argument components and relation annotations between the annotators. In addition, I calculated some statistics of the final corpus.

To evaluate the reliability of the argument component and argumentative relation annotations, I followed the approach of Stab and Gurevych (2014a).

**Inter-Annotator Agreement Argument Components**

With regard to the argument components, two strategies were used. Since there were no predefined markables, annotators not only had to identify the *type of argument component*, but also its *boundaries*. In order to assess the latter, I use Krippendorff's $\alpha_U$ (Krippendorff, 2004), which allows for assessing the reliability of an annotated corpus, considering the differences in the markable boundaries. To evaluate the annotators' agreement in terms of the selected category of an argument component for a given sentence, I calculate percentage agreement and two chance-corrected measures, multi-$\pi$ (Fleiss, 1971) and Krippendorff's $\alpha$ (Krippendorf, 1980). I decided to operate at sentence level, since only 20.56% of the sentences in the corpus contain annotations of different argument components. Thus, evaluating the reliability at sentence level served as a good approximation of the inter-annotator agreement. At the level of individual sentences, 33.45% contain a claim, 35.64% a premise and 55.52% a non-annotation. 20.56% of the sentences contain several annotations, with a combination of a premise and a non-argumentative span (32.21%) and a combination of a claim and a non-argumentative span (35.78%) representing the majority of those cases. Only 4.08% contain both a premise and a claim. At the token level, the following class distribution is achieved: 43.54% claim, 45.42% premise and 11.04% are not annotated.

Table 15 displays the resulting inter-annotator agreement scores. I obtained an inter-

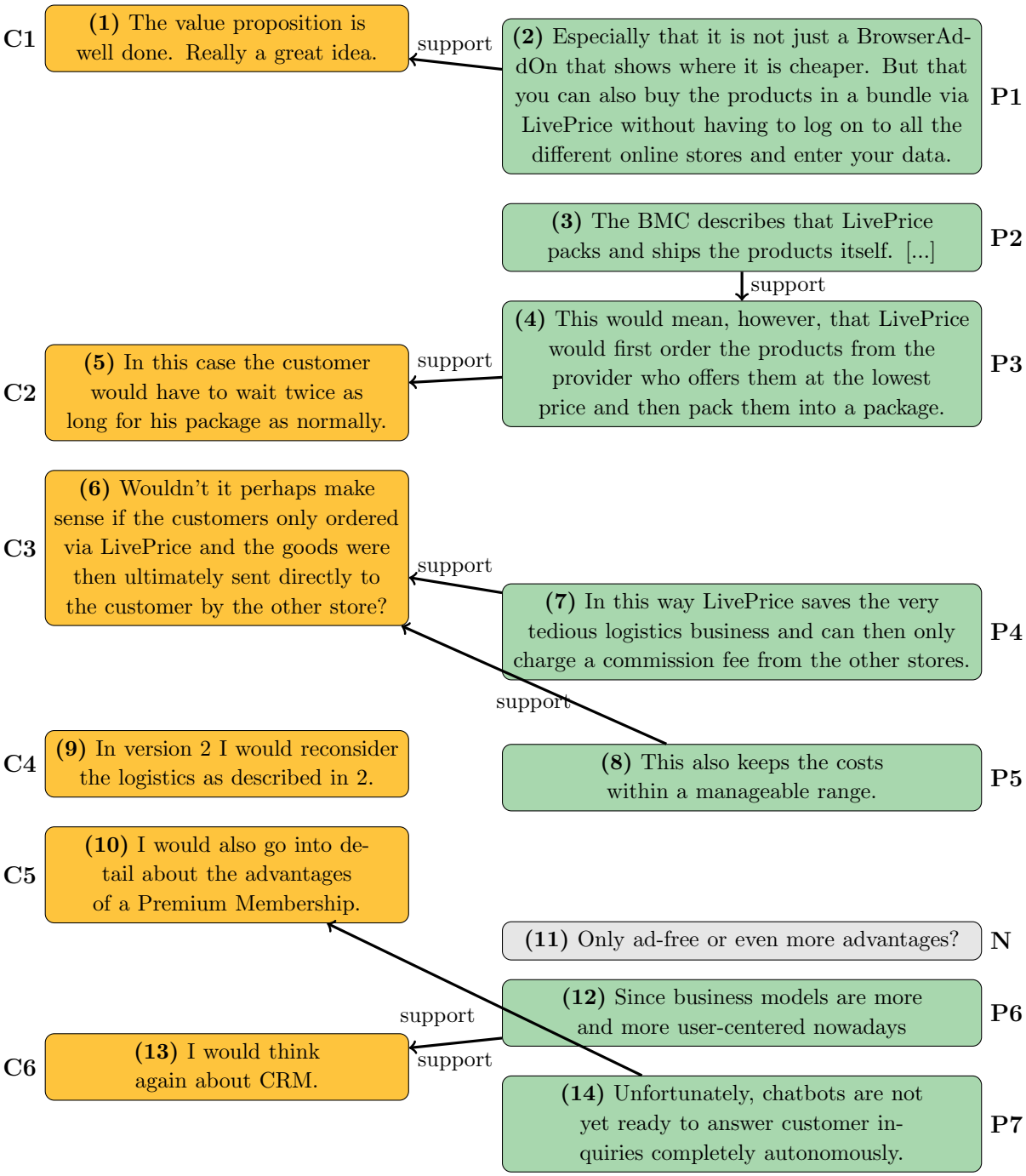



Figure 36: Fully annotated example of a peer review according to the annotation scheme and guidelines. The left column represents the claims (C1 - C6), while the premises are listed on the right (P1 - P7). N signifies a non-argumentative text unit.
Source: Own illustration

| | % | Multi-$\pi$ | Krippendorff's $\alpha$ | Krippendorff's $\alpha_U$ |
|---|---|---|---|---|
| **Claim** | 0.7865 | 0.5547 | 0.5549 | 0.4404 |
| **Premise** | 0.7663 | 0.5106 | 0.5108 | 0.4776 |
| **None** | 0.8572 | 0.6410 | 0.6412 | 0.3590 |

Table 15: Inter-rater agreement of argument component annotations.
Source: Own illustration

annotator agreement of 78.65% for the claims and 76.63% for the premises. The corresponding multi-$\pi$ scores are 55.47% and 51.06%. Regarding Krippendorff's $\alpha$, a score of 55.49% and 51.08% is obtained, indicating a moderate agreement for both categories. With a score of 44.04% and 47.76%, the unitized $\alpha$ of both the claim and premise annotations is slightly smaller compared to the sentence-level agreement. Thus, the boundaries of argument components are less precisely identified in comparison to the classification into argument types. Yet the scores still suggest that there is a moderate level of agreement between the annotators. With a score of $\alpha_U$=35.90%, the boundaries of non-argumentative units are less reliably detected. In contrast, the agreement scores targeting the component types are considerably higher for the non-argumentative spans as compared to the claims and premises (85.72%, 64.10%, 64.12%), indicating a substantial agreement between the annotators. Hence, I conclude that the annotation of the argument components in student-generated peer reviews is reliably possible.

**Inter-Annotator Agreement Argumentative Relations**

| | % | Multi-$\pi$ | Krippendorff's $\alpha$ |
|---|---|---|---|
| **Support** | 0.9413 | 0.4903 | 0.4903 |
| **Non-support** | 0.9413 | 0.4903 | 0.4903 |

Table 16: Inter-rater agreement of argumentative relation annotations.
Source: Own illustration

To evaluate the reliability of the argumentative relations, I used the set of all pairs between argument components that were possible during the annotation task according to the annotation scheme, i.e. all pairs between a claim and a premise and between two premises. In total, the markables include 4,792 pairs of which 7.41% are annotated as support relations and 0.35% as attack relations. 92.24% of the possible pairs were not identified by an annotator. Since the number of attack relations is so small, I decided to focus on the support relations, distinguishing only between the two types *support* and *non-support*. I obtained an inter-annotator agreement of 94.13% for both support and non-support relations. The corresponding multi-$\pi$ and Krippendorff's $\alpha$ scores both amount to 49.03% (see Table 16). Therefore, I conclude that argumentative relations, can be reliably annotated in student-generated argumentative peer reviews.

#### 5.2.2.5 Corpus Statistics

| | total | mean | standard deviation | min | max | median |
|---|---|---|---|---|---|---|
| **Sentences** | 20,217 | 20.217 | 9.4299 | 2 | 69 | 19.0 |
| **Tokens** | 246,980 | 246.98 | 101.4358 | 40 | 720 | 236.0 |

Table 17: Distribution of *sentences* and *tokens* in the created corpus. Mean, standard deviation, minimum, maximum and median refer to the number of sentences and tokens, respectively, per document.
Source: Own illustration

Table 17, Table 18, Table 19 and Table 20 present some statistics of the final corpus. It consists of 1,000 student-written peer reviews in German that are composed of 20,125 sentences with 246,980 tokens in total. Hence, on average, each document has 20 sentences and 272 tokens. A total of 7,996 claims (31.64%) and 8,479 premises (33.55%) were annotated. 8,796 textual spans (34.81%) were not identified as an argument component ("None"). On top of that, 8,291 support (97.03%) and 254 attack (2.97%) relationships were marked up by the annotators. The little percentage of attack relations is explainable due to the domain of the annotated texts. The nature of peer reviews is to provide feedback about a certain topic by highlighting strengths, weaknesses and suggesting improvements. Usually, students used premises to support their claims, in fact, only about 3% of the claims were further elaborated by discussing them more controversially through attacking premises.

| | total | mean | standard deviation | min | max | median | percent |
|---|---|---|---|---|---|---|---|
| **Claim** | 7,996 | 7.996 | 3.6487 | 1 | 25 | 7.0 | 0.3164 |
| **Premise** | 8,479 | 8.479 | 4.8346 | 0 | 26 | 8.0 | 0.3355 |
| **None** | 8,796 | 8.796 | 3.9545 | 2 | 26 | 8.0 | 0.3481 |
| **All** | 25,271 | 8.4237 | 4.1878 | 0 | 26 | 8.0 | 1.0 |

Table 18: Distribution of types of *argument components* in the created corpus.
Source: Own illustration

| | total | mean | standard deviation | min | max | median | percent |
|---|---|---|---|---|---|---|---|
| **Support** | 8,291 | 8.291 | 4.7771 | 0 | 26 | 7.0 | 0.9703 |
| **Attack** | 254 | 0.254 | 0.6196 | 0 | 5 | 0.0 | 0.0297 |
| **All** | 8545 | 4.2725 | 5.2681 | 0 | 26 | 2.0 | 1.0 |

Table 19: Distribution of types of *argumentative relations* in the created corpus.
Source: Own illustration

Table 20 presents the distribution of support relationships in the corpus. With 46.30%, the majority of claims is not supported by any premise. 39.07% of the claims are backed up by exactly one premise, while 10.93% are supported by two premises. Patterns with

more than two supporting premises rarely occur in the dataset. The majority of premises is not backed up by another premise (75.07%). However, in 21.08% of the cases, there is one supporting premise. In that way, more complex reasoning chains are established.

| | #0 support | #1 support | #2 support | #3 support | ≥ #4 support |
|---|---|---|---|---|---|
| **Claim** | 0.4630 | 0.3907 | 0.1093 | 0.0230 | 0.014 |
| **Premise** | 0.7507 | 0.2108 | 0.0289 | 0.0073 | 0.0024 |

Table 20: Distribution of patterns of *support relationships* in the created corpus. The columns denote the percentage of claims and premises, respectively, that are supported by the specified number of premises.
Source: Own illustration

#### 5.2.2.6 Conclusion

In this section, I introduced an argumentation annotation scheme as well as an annotated corpus of persuasive student-written peer reviews extracted from a real-world learning scenario that can be leveraged to give students feedback on the quality of their argumentation (e.g., by highlighting insufficiently backed claims) through an AI-based argumentation writing support system.

The contribution is threefold: (1) I derive an annotation scheme for a new data domain for argumentation mining based on argumentation theory and previous work on annotation schemes for persuasive student essays (Stab and Gurevych, 2014b, 2017); (2) I present an annotation study based on 50 persuasive peer reviews that were annotated by three native German speakers, demonstrating that the annotation of student-generated peer reviews is reliably possible; and (3) I present the final and freely available corpus of 1,000 student peer reviews collected in the lecture about business innovation in German language. The corpus as well as the underlying annotation scheme serves as the underlying dataset for the implementation of AI-based argumentation writing support systems in section 5 and for conversational argumentation learning systems in section 6 of this dissertation.

With this research, I hope to encourage fellow researchers to leverage the annotation scheme and argumentation corpus to design and develop writing support systems for students in large-scale learning scenarios.

After the annotation of a novel corpus for AI-based argumentation writing support systems, the next chapter incorporates the demonstration and evaluation of the proposed architecture of AI-based argumentation writing support systems. Therefore, the human–computer interaction of the instantiated argumentation writing support system ArgueLearn is explained. The developed corpus from this section is used to model argumentative

discourse in student-written peer reviews and, therefore, provides the intelligent back end of the learning tool.

### 5.2.3 Instantiation of the AI-based Argumentation Writing Support System ArgueLearn

Following the above-mentioned design principles, ArgueLearn is built as a responsive web-based application that can be used on all kinds of devices. The front end of ArgueLearn was developed with recent web technologies, including HTML5, Cascading Style Sheets (CSS), and JavaScript (JS). A user can access ArgueLearn from any web-enabled device regardless of screen size or operating system and can log in with a single sign-on for easy access (F1). It provides the user with a simple and intuitive text input field and a word count (F2) in which they can write or copy a text. Next to the input field, the users receive feedback on their text's argumentation structure on a personal learning dashboard (F3 – F6). A screenshot of ArgueLearn and its different functionalities (e.g., *F1 - F6*) can be seen in Figure 39. In total three different versions of ArgueLearn have been designed and implemented in the course of my research. The early design concept of ArgueLearn can be seen in Figure 37. The second version of ArgueLearn, which was used for the empirical evaluation in study 1, is illustrated in Figure 38.

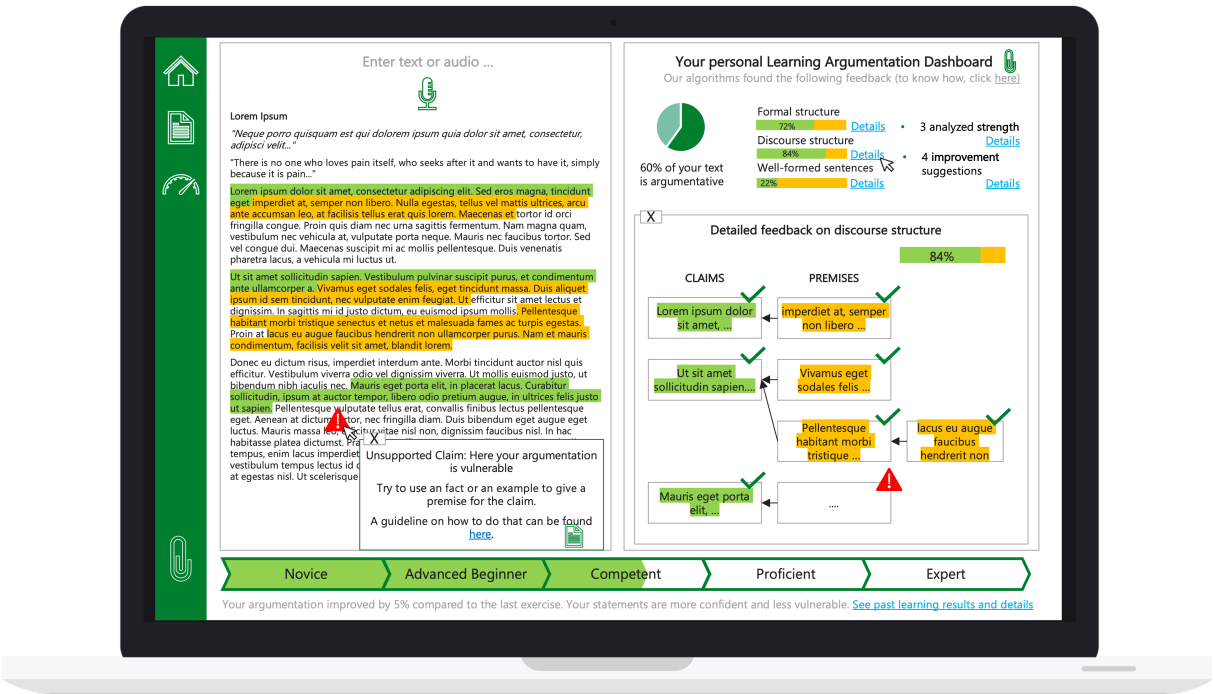



Figure 37: Early design concept of ArgueLean for receiving initial user feedback in the pre-study.
Source: Own illustration

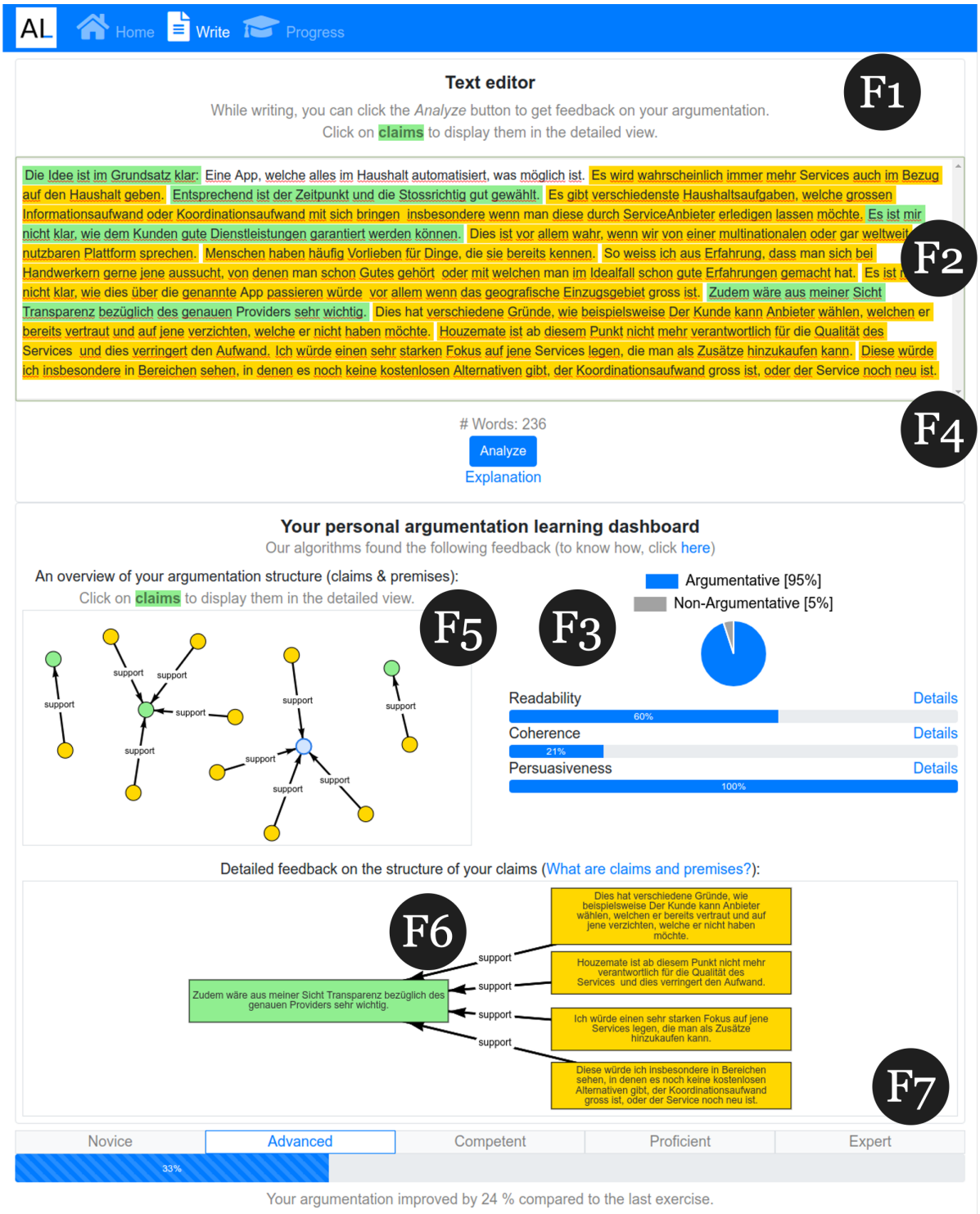



Figure 38: Screenshot of the second version of ArgueLearn's interface with instantiated functionalities (F1 to F7). This version of ArgueLearn was used for evaluating the design principles in a laboratory experiment (study 1).
Source: Own illustration

The dashboard provides feedback on different granularity levels, allowing users to control the amount of feedback information required. A visual graph-based representation

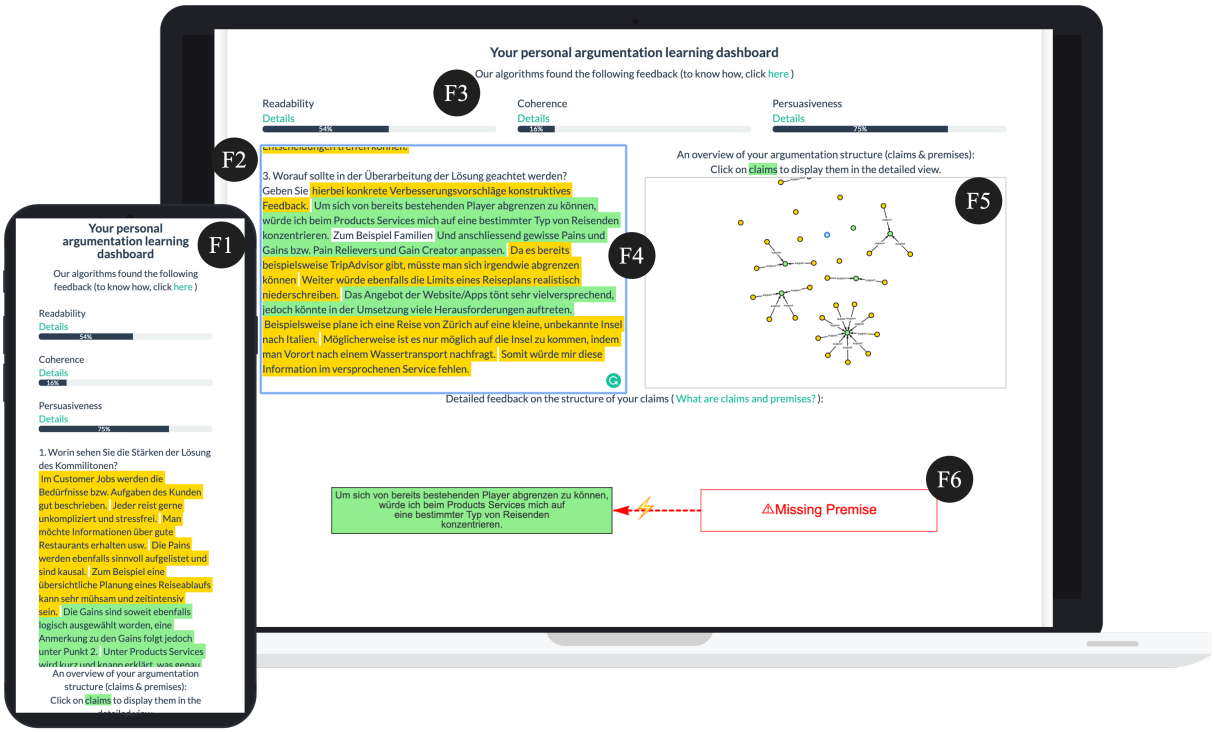



Figure 39: Screenshot of the final version of ArgueLearn's interface with instantiated functionalities (F1 to F6). This version was instantiated and evaluated in a field experiment (study 2).
Source: Own illustration

of the submitted text's argumentation structure (F5) and three summarizing scores give an initial overview of the text's quality (F3). The identified claims are green and the premises are highlighted in yellow to provide users with instant feedback on their own submitted input (F4). By clicking on the marked text fields or the nodes in the graph, a more detailed view of the discourse of the argument will appear (F6). This indicates whether a claim is sufficiently supported (see Figure 39) or if it lacks a premise (as in F6). This function provides learners with clear steps on how to improve their texts' persuasiveness and formal quality. Moreover, best practices and explanations about argumentation and argumentation theory are provided by clicking the "explanation" or "help" button, providing the user with an orientation and context to improve their writing quality. The three summarizing scores—readability, coherence, and persuasiveness (F3)—provide the users with a ranking of their text for superficial instant feedback. By clicking on the scores or on "details," the methodology for calculating the scores as well as concrete hints and explanations of how the learner can increase their score level will be shown (see Figure 40). The scores are calculated using the following metrics: readability is defined as how readable the text is based on the Flesch reading ease score (Flesch, 1943); coherence indicates the proportion of sentences that are connected via discourse markers; and persuasiveness concerns the proportion of claims that are supported by premises compared

to unsupported claims. Moreover, ArgueLearn provides a learning progress bar, which actively monitors the learner's past and current learning development to convey the goal and purpose of learning (F7, see Figure 38), and an introduction to argumentation theory based on the Toulmin model (Toulmin, 2003).

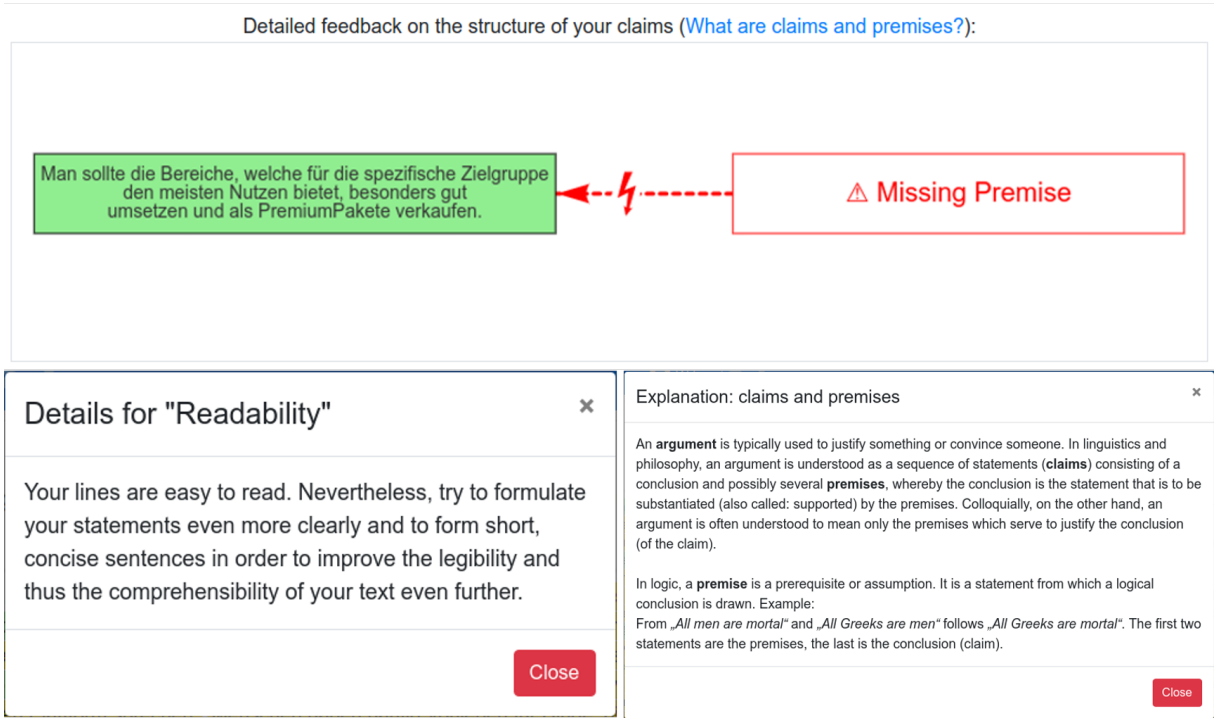



Figure 40: Top: screenshot of detailed discourse feedback of unsupported claim. Below: screenshot of exemplary explanations and details of ArgueLearn.
Source: Own illustration

### 5.2.4 Argumentation Mining Algorithm of ArgueLearn

To design an individual and adaptive feedback tool, I collected and annotated an own corpus to train and tune a model that fulfils the users' requirements to give instant feedback on their texts. The corpus collection and annotation is described in the previous section.

To provide students with feedback on the argumentation quality of their texts, I, first of all, implemented an approach for detecting arguments in them. This approach consists of two subtasks. In the first step, I identified the components of arguments in terms of claims and premises. Next, I determined whether there is an argumentative relation between a pair of argument components. To do so, I followed the approach described in Stab and Gurevych (2017), a state of the art approach for identifying argumentative discourse structures in persuasive essays.

| Feature | Description |
|---|---|
| Unigrams | Binary and lemmatized unigrams of the component and its preceding token |
| Dependency Tuple | Lemmatized dependency tuples (2k most frequent) |
| Token statistics | Number of tokens of component, covering paragraph and covering sentence; number of tokens preceding and following the component in its sentence; ratio of component and sentence tokens |
| Component position | Component is first or last in paragraph; component present in introduction or conclusion; Relative position in paragraph; number of preceding and following components in paragraph |
| Type indicators | Forward, backward, thesis or rebuttal indicators present in the component or its preceding tokens |
| First-person indicators | "I", "me", "my", "mine", or "myself" present in component or its preceding tokens |
| Type indicators in context | Forward, backward, thesis or rebuttal indicators preceding or following the component in its paragraph |
| Shared phrases | Shared noun phrases or verb phrases with the introduction or conclusion (number and binary) |
| Subclauses | Number of subclauses in the covering sentence |
| Depth of parse tree | Depth of the parse tree of the covering sentence |
| Tense of main verb | Tense of the main verb of the component |
| Modal verbs | Modal verbs present in the component |
| POS distribution | POS distribution of the component |
| Type probability | Conditional probability of the component being a major claim, claim or premise, given its preceding tokens |
| Discourse Triples | PDTB-discourse relations overlapping with the current component |
| Combined word embeddings | Sum of the word vectors of each word of the component and its preceding tokens |

Table 21: Features used for argument component identification according to Stab and Gurevych (2017) and the NLP feature taxonomy from subsection 4.1.
Source: Stab and Gurevych (2017)

#### 5.2.4.1 Subtask 1: Argument Component Identification

The identification of argument components is considered as a sentence-level multi-class classification task, where each sentence in the dataset is labeled as either *claim*, *premise* or *non-argumentative*. Hence, apart from classifying argument components as claims or premises, this task includes the separation of argumentative from non-argumentative text units. To ensure an equal distribution of classes among training and test sets in the experiments, I performed a stratified split of the data set into a 80% training set and a 20% test set, resulting in the distribution of 32% claims, 32% premises and 36% non-argumentative spans (for both training and test set). I used several classifiers (Support Vector Machine (SVM), Logistic Regression, Random Forest, Multinomial Naive Bayes (NB), Gaussian NB, Nearest Neighbor, and AdaBoosted Decision Tree) for the task of argument component identification. To tune the parameters of the models, I applied grid search. For pre-processing the documents, I used the spacy parser.[10] The features given in Table 21 were extracted for training a model to perform the task of argument component identification following the taxonomy explained in subsection 4.1. I found that an SVM achieves the best results, with an accuracy of 65.4% on the test set. The resulting argument structure is visualized directly in the student-generated text by highlighting claim components in green and premise components in yellow, while non-argumentative text spans are not marked up.

#### 5.2.4.2 Subtask 2: Argument Relation Identification

The identification of argumentative relations is considered a binary classification task of argument component pairs, where each pair is classified as either *support* or *non-support*. All possible combinations of argument components are tested. Like before, I randomly split the dataset in a 80% training set and a 20% test set and determined the best performing system using 10-fold cross-validation on the training set. I used the same pre-processing pipeline as described in the previous paragraph and extracted the features detailed in Table 22. The comparison of several classifiers (SVM, Logistic Regression, Random Forest, Multinomial NB, Gaussian NB, Nearest Neighbor and AdaBoosted Decision Tree) revealed that an SVM achieves the best results for the corpus, obtaining an accuracy of 72.1% on the test set. To tune the parameters of the model, I again used grid search. This is satisfying compared to other studies on student-written argumentation identification. For example, Stab and Gurevych (2017) reached an f1-score of 73% for argumentation stance classification of student-written texts in English.

[10] https://spacy.io/

| Feature | Description |
|---|---|
| Unigrams | Binary lemmatized unigrams of the source and target components including preceding tokens (500 most frequent) |
| Part-of-speech | Binary POS features of source and target components |
| Production rules | Production rules extracted from the constituent parse tree (500 most frequent) |
| Token statistics | Number of tokens of source and target |
| Component statistics | Number of components between source and target; number of components in covering paragraph |
| Position features | Source and target present in same sentence; target present before source; source and target are first or last component in paragraph; pair present in introduction or conclusion |
| Indicator source/target | Indicator type present in source or target |
| Indicators between | Indicator type present between source or target |
| Indicators context | Indicator type follows or precedes source or target in the covering paragraph of the pair |
| Discourse Triples | Binary discourse triples of source and target |
| Pointwise mutual information | Ratio of tokens positively or negatively associated with incoming or outgoing relations; Presence of words negatively or positively associated with incoming or outgoing relations |
| Shared nouns | Shared nouns between source and target components (number and binary) |

Table 22: Features used for argument relation identification the NLP feature taxonomy from subsection 4.1.
Source: Stab and Gurevych (2017)

Finally, I implemented the argumentation mining algorithm as an intelligent back end into our prototypical instantiation of ArgueLearn. The resulting argumentative discourse structures are visualized in terms of a directed graph connecting a claim with its supporting premises (see Figure 39, F6). Unsupported claims are highlighted to indicate that further support is warranted (see Figure 39, F6).

Besides, I calculated a number of summary scores for providing students with an overview of the quality of their argumentation based on previously extracted argumentative discourse structures, including

- **Readability**: How readable is the text based on the Flesch Reading Ease score (Flesch, 1943)?
- **Coherence**: How large is the proportion of sentences that are connected via discourse markers?

- **Persuasiveness**: How large is the proportion of claims that are supported by premises as compared to unsupported claims?

## 5.3 Demonstration and Evaluation of an AI-based Argumentation Writing Support System

In this section, I demonstrate the application and evaluation of the design principles for AI-based argumentation writing support systems in the context of higher education. I present a formal evaluation of the proposed design principles as part of the artifacts of the AI-based argumentation writing support system, with two experimental studies (study 1 and study 2).

The design science research methodology comprises two main phases: demonstration and evaluation (Peffers et al., 2007). The demonstration displays the artifacts' ability to solve the identified research problem (Venable et al., 2016). It can be described as a proof-of-concept evaluation aiming to demonstrate that the artifact can achieve its design purpose in a particular context. The evaluation aims to formally validate the research artifacts' effectiveness in addressing the problem (Peffers et al., 2007; Venable et al., 2016). Hence, I performed an artificial ex-post evaluation of ArgueLearn in study 1 to evaluate the proof of concept and a naturalistic ex-post evaluation of ArgueLearn in study 2 to evaluate the proof of value (Venable et al., 2016). Accordingly, it is important to evaluate the proof of an artifact's value to verify whether the design principles and the resulting artifact are valuable in real-life settings.

Study 1 was conducted in a university lab designed for behavioral studies under controlled circumstances. Study 2 was conducted in a large-scale master lecture at the University of St. Gallen in Switzerland. Thus, I could ensure that no participant took part in both studies. Before the two empirical studies, I performed several small-scale evaluations of the distinct artifacts (e.g., testing the initial human—computer interaction concept of ArgueLearn presented in Figure 37). My aim was to conduct iterative evaluations to ensure that the instantiated design functionalities correctly addressed the design principles and that the technical artifacts (artifact 1, 2 and 3) enabled the learners to receive individual self-evaluations irrespective of instructor, time, and place. Table 23 presents an overview of the demonstration and evaluation of ArgueLearn.

### 5.3.1 Pre-Study: Design Evaluation

Based on the requirements and design principles, I designed clickable mock-ups of different AI-based argumentation writing support systems. For the evaluation, I followed an ex-ante evaluation using an artificial evaluation setup, as proposed by Venable et al. (2016). Based on Venable's criteria, the pre-study design evaluation's objective was to:

| Study | Objective/ guiding question | Result |
|---|---|---|
| **Pre-Study:** Technical- and design evaluations of the artifacts | Ensure all design principles are correctly addressed and that technical artifacts provide high-quality argumentation self-evaluation. | Satisfying accuracy and performance of an algorithm based on tuned features; positive user-feedback on the implementation of the design principles into my instantiation |
| **Study 1:** Proof of concept evaluation in the laboratory | *Does an AI-based argumentation writing support system based on my design principles help learners improve their argumentation skills in short-term scenarios?* | Individuals using ArgueLearn wrote texts with better formal quality of argumentation as well as better perceived quality of argumentation compared to those using the scripting benchmark approach. |
| **Study 2:** Proof of value evaluation in a real-life setting over three months | *To what extent does an AI-based argumentation writing support system based on my design principles influence learners' long-term metacognitive argumentation skills in large-scale learning settings?* | Individual feedback on learners' argumentation errors on business model feedback helps them to write more convincing texts in another argumentation domain (reviews about social debates). Learners using ArgueLearn wrote texts with a better formal quality of argumentation than those receiving only general or no argumentative writing support. |

Table 23: Overview of demonstration and evaluation of the AI-based argumentation writing support system guided by Venable et al. (2016).
Source: Own illustration

- Verify the utility and value of the artifact to achieve its stated purpose.
- Identify weaknesses and areas for improvement in the artifact design.
- identify change requests from students, side effects, or undesirable consequences of its use.

To reach this goal, I conducted four evaluation series with 46 different users (around 12 users per series). I performed different qualitative assessments involving observation, participant feedback, and unstructured interviews (Tuunanen and Peffers, 2018). These users differed from those recruited for studies 1 and 2 as well as those from the semi-structured interviews but were also students from the university with a similar age and gender distribution. I started the testing with three low-fidelity paper prototypes and later with two digital mock-ups of ArgueLearn. For example, I hypothesized that I can instantiate the leaner control function (DP2, see prior subsection) with a graph-based representation of their discourse level. I tested this with a paper prototype that provided

visual feedback on the argumentation of a given text (simulating the feedback algorithm by a human). The hypothesis was validated with 10 users who overall liked the concept of graph-based feedback. Therefore, the final prototypes contribute to that with a graph engine that provides a visual representation of the argumentation of an analyzed text.

### 5.3.2 Study 1: Proof of Concept Evaluation in a Laboratory Experiment

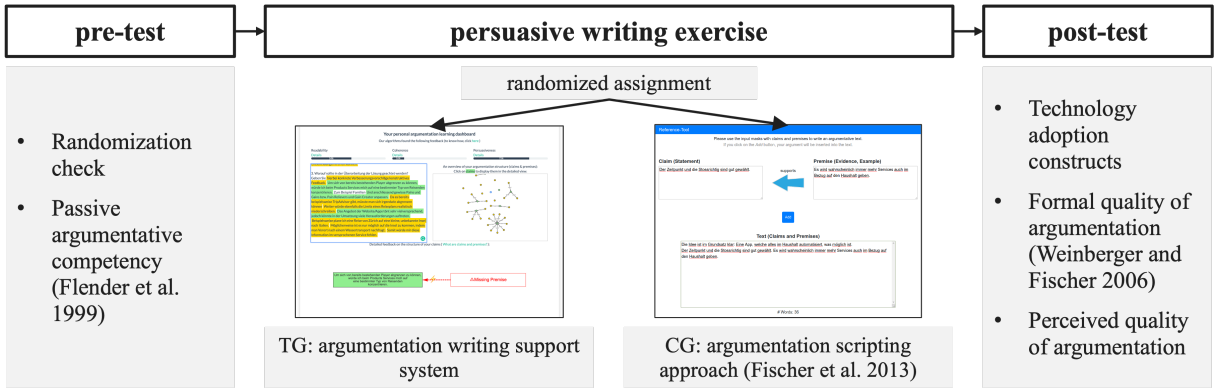



Figure 41: Overview of experimental procedure of study 1 to evaluate if an AI-based argumentation writing support system based on my design principles improves persuasive writing more so than traditional argumentation scripting.
Source: Own illustration

After evaluating the technical artifacts and the design in a pre-study, I performed a proof of concept evaluation as the first demonstration of my design principles (Peffers et al., 2007; Venable et al., 2016). The objective was to verify the artifacts' ability to solve the identified research problem (Venable et al., 2016). I conducted study 1 as a laboratory experiment in which participants were asked to provide peer feedback on an essay. The treatment group used ArgueLearn, while participants in the control group used a discussion scripting application, in accordance with Fischer et al. (2013) as a well-cited and empirically proven benchmark to foster the formal quality of argumentation. I recruited students from the university through social networks and mailing lists to participate. After randomization, I had 24 participants in the treatment group and 30 in the control group. I invited them to the laboratory of the university, where I conducted the study on identical devices. Participants in the treatment group had an average age of 23.8 (SD= 3.86); 15 were male, and 9 were female. In the control group, the participants' average age was 23.03 (SD= 2.12); 22 were male and 8 female. All participants were compensated with an equivalent of about 15 USD for the 30–40-minute experiment.

#### 5.3.2.1 Alternative Learning System: Discussion Scripting

To evaluate ArgueLearn, I compared it to a discussion scripting application, which I also built myself. Implementing an own discussion scripting approach allowed to control the differences and similarities in the design between the discussion scripting tool and ArgueLearn. For the design, I followed the approach of Stegmann et al. (2012), since it is well-cited and empirically proven to foster the formal quality of argumentation of students. The learning tool supports the writing process of users with input masks (see Figure 42, *F1* and *F2*). Users can use these input masks to compose a formally correct argument. By clicking "*Add*", the argument is attached to the text field (*F3*). If the user decides to not use the input masks, they can also write directly into the final text field. To keep ArgueLearn and the discussion scripting approach consistent with each other, there are many functions that are shared between them. First, the introduction text is the same across both apps. The help and explanation buttons in the discussion scripting approach correspond respectively to the same buttons in ArgueLearn. Moreover, both text fields consist of a word count function to provide guidance in the writing process. I used the design of the particular input mask of this scripting approach and tested it in the non-collaborative experimental use case to measure the influence of the approach on the argumentation quality of students' texts. However, both approaches, the adaptive system, and the scripting approach could be used in a collaborative learning scenario, e.g., where students give each other feedback on a business model and discuss these.

#### 5.3.2.2 Experimental Design

The experiment consisted of three main phases: (1) pre-test phase, (2) individual writing phase and (3) post-test phase. The pre- and post-phases were consistent for all participants. In the writing phase, the treatment group used ArgueLearn and the control group wrote a text using the alternative tool.

**(1) Pre-test phase:** The experiment started with a pre-survey with 14 questions. Here, I tested three different constructs to assess whether the randomization resulted in randomized groups. First, I asked four items to test the personal innovativeness in the domain of information technology of the participants following Agarwal and Karahanna (2000). The items were "I like to experiment with new information technologies", "If I heard about a new information technology, I would look for ways to experiment with it, "In general, I am hesitant to try out new information technologies" and "Among my peers, I am usually the first to try out new information technologies". Second, I tested the construct of feedback-seeking of individuals following Ashford (1986). Example items are: "*It is important for me to receive feedback on my performance.*" or "*I find feedback on my performance useful.*" Both constructs were measured with a 1- to 5-point Likert scale (1:

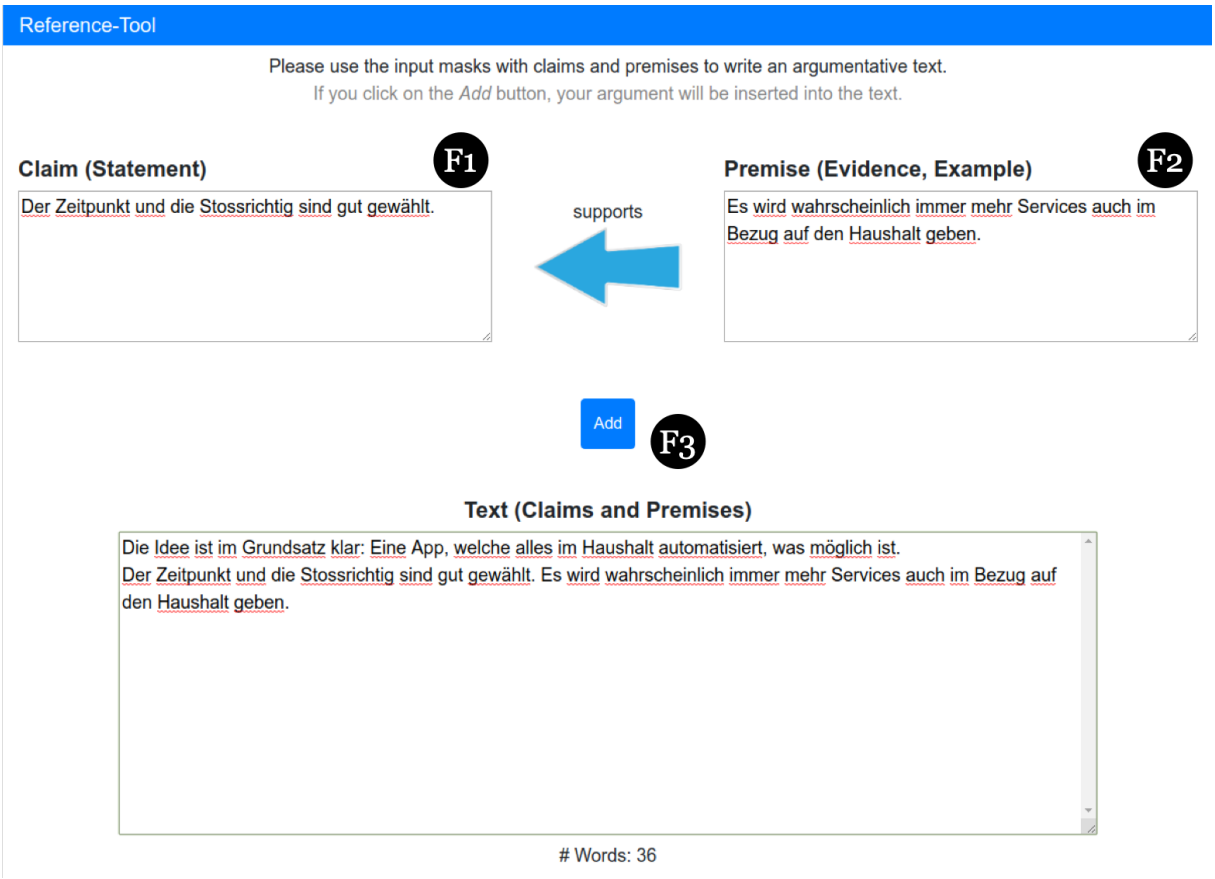



Figure 42: Screenshot of discussion scripting approach: a user enters a claim and the corresponding premise into the input masks and adds it to the final text.
Source: Own illustration

totally agree to 5: totally disagree, with 3 being a neutral statement). Third, I captured the construct of passive argumentative competency following the design of Flender et al. (1999), since it is a proven construct to measure argumentative competencies in German. I wanted to control for the argumentative competencies, since I later measured the formal and perceived quality of argumentation of the written texts. Participants were asked to read a discussion of two teachers concerning the topic "*Does TV make students aggressive?*" I retrieved the topic with the discussion as well as the measurements from (Flender et al., 1999). Based on the discussion, I asked the participants three questions concerning the argumentation structure and the content of the text with multiple choice answers: "*What kind of argumentation style or structure is used*?", "*How can a new argument be added to the discussion?*" and "*Which of the following standpoints do both parties agree on*?" (Flender et al., 1999). Additionally, participants were asked how sure they were about the answers on a 1- to 5-point Likert scale (1: very sure, 5: not very sure, with 3 being a neutral statement). The competencies were then measured with a certain score from 0 to 27 following the measurements of Flender et al. (1999).

**(2) Writing phase:** In the writing phase of the experiments, I asked the participants to write a review about the argumentation of both parties (pro and contra) concerning the weaknesses and strengths of their argumentation. The participants were told to spend at least 15 minutes on writing this review. A countdown indicated them the remaining time. They were only able to continue the experiment after the countdown was finished. The treatment group was using ArgueLearn to write the review, the control group was using the reference tool. I did not provide any introduction to any of the tools. The students using ArgueLearn retrieved individual and adaptive feedback based on the feedback algorithms. Participants using the reference tool retrieved help based on input formats during the writing process.

**(3) Post-test phase:** In the post-survey, I measured perceived usefulness, intention to use and ease of use following the technology acceptance model of Venkatesh et al. (2003); Venkatesh and Bala (2008) and captured the demographics. In total, I asked 16 questions. Example items for the three constructs are: "*Imagine the tool would be available in your next course, would you use it?*", "*The use of the argumentation tool enables me to write better argumentative texts.*" or "*I would find the tool to be flexible to interact with*". Moreover, I were asking three qualitative questions: "*What did you particularly like about the use of the argumentation tool?*", "*What else could be improved?*" and "*Do you have any other ideas?*"

### Measurement of Argumentation Quality

Besides measuring the technology acceptance, the main objective was to measure the

quality of the written texts from both groups to evaluate the main hypothesis. Therefore, I measured two main variables: (1) the formal quality of argumentation and (2) the perceived quality of argumentation.

**(1) Formal quality of argumentation**: The written peer reviews were analyzed for the formal quality of argumentation. I applied the annotation scheme for argumentative knowledge construction described by Weinberger and Fischer (2006). This annotation scheme has been applied in various studies and is proven to have high objectivity, reliability, and validity (e.g., Stegmann et al. (2012)). To measure the argumentation's formal quality, the annotator had to distinguish between (a) unsupported claims, (b) supported claims, (c) limited claims, and (d) supported and limited claims. A more precise description of the scheme can be found in Weinberger and Fischer (2006). Two annotators, who had already participated in the annotation process for the corpus, annotated the participants' texts based on the annotation guidelines and prior experience. The formal quality of individual users' argumentation was then defined by the number of arguments that the user wrote. Following Stegmann et al. (2012), only supported, limited, and supported and limited claims were counted as argumentation.

**(2) Perceived quality of argumentation:** The perceived quality of argumentation was annotated by two different annotators. The objective was to subjectively judge how persuasive the given argumentation was on a 5-point Likert scale (1: not very persuasive, 5: very persuasive). I took the mean of both annotators as a final variable for the texts' perceived and formal quality of argumentation.

#### 5.3.2.3 Results

To demonstrate my design principles, my aim was to evaluate if an AI-based argumentation writing support system based on the design principles helps individuals to write more argumentative more convincing texts. To guide the proof of concept evaluation, I aim to answer two research questions (RQ):

**RQ1:** *Do students perceive ArgueLearn to be useful and easy to use, and would they continue to use it in the future?*

**RQ2:** *How effective is ArgueLearn with helping users to write more persuasive texts compared to the traditional, proven approach?*

The first research question was answered by comparing the constructs of perceived usefulness, intention to use, and ease of use for participants using ArgueLearn compared to

participants using the alternative tool. In particular, I used a double-sided t-test to evaluate whether the means of the constructs are significantly different. Moreover, I compared the results of ArgueLearn to the midpoints scale to validate a general positive technology acceptance as done in Lehmann et al. (2016).

To evaluate the second research question, I compared the formal quality of argumentation as well as the perceived quality of argumentation between the written text of the treatment and the control group. I performed a double-sided t-test to assess whether the differences between both groups are statistically significant.

To control for potential effects of interfering variables with the small sample size and to ensure that randomization was successful, I compared the differences in the means of the three constructs included in the pre-test. For all three constructs, including personal innovativeness, feedback-seeking of individuals, and passive argumentative competency, I obtained p values greater than 0.05 between the treatment and the control group. The p value for personal innovativeness between both groups was $p = 0.801$; for feedback-seeking of individuals $p = 0.624$; and for passive argumentative competency $p = 0.375$, verifying that no significant difference exists between the groups.

| Group | Intention to use | Perceived usefulness | Perceived ease of use |
|---|---|---|---|
| **Mean ArgueLearn** | 2.33 | 2.52 | 2.17 |
| **Mean reference tool** | 3.5 | 3.28 | 2.84 |
| **SD ArgueLearn** | 0.59 | 0.58 | 0.65 |
| **SD reference tool** | 1.14 | 1.12 | 1.08 |
| **p-value** | <0.001 | 0.006 | 0.012 |

Table 24: Results of the technology acceptance of ArgueLearn and the reference tool on a 1 - 5 Likert Scale (1: high, 5: low).
Source: Own illustration

**Technology Acceptance**

For the technology acceptance, I calculated the average of every construct. The answers were provided on a 1- to 5-point Likert scale (1: very sure, 5: not very sure).

First, I compared the results of ArgueLearn with the results of the alternative tool. The perceived usefulness of ArgueLearn was rated with a mean value of 2.52 (SD= 0.58) and the average of perceived use of ease of ArgueLearn was 2.17 (SD= 0.65). The mean value of intention to use of participants using ArgueLearn as a writing tool was 2.33 (SD= 0.58). These values are significantly better than the results of the alternative scripting approach. For perceived usefulness, I observed a mean value of 3.2 (SD= 1.12) and for perceived ease of use the value was 2.83 (SD= 1.08) for participants from the control group. The mean value for the intention to use was 3.5 (SD= 1.13).

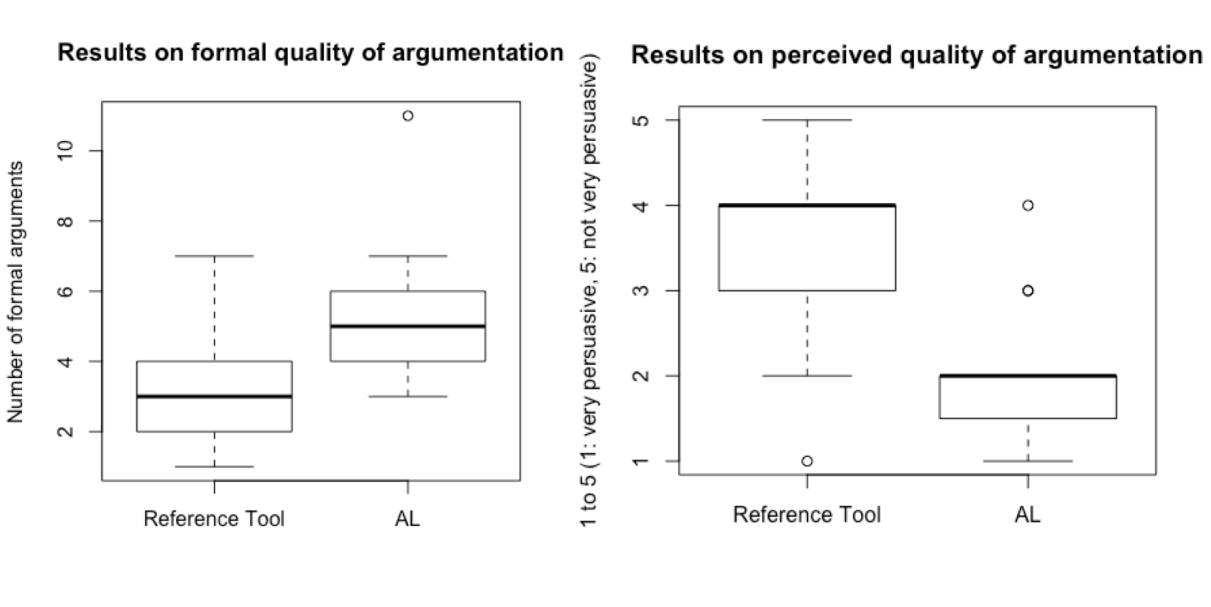



Figure 43: Results on formal (left) and perceived (right) quality of argumentation between both tools.
Source: Own illustration

The results clearly show that the participants of the experiment rated the acceptance of ArgueLearn as an adaptive feedback tool positively compared to the usage of the alternative application. The statistical significance was also proven in a double-sided t-test for all three constructs (see Table 24). Moreover, the mean values of ArgueLearn are also very promising when comparing the results to the midpoints. All results are better than the neutral value of 3. Especially the perceived usefulness for writing better argumentative texts and the intention to use ArgueLearn as a writing support tool show promising results. A positive technology acceptance is especially important for learning tools to ensure students are perceiving the usage of the tool as helpful, useful, and easy to interact with. This will foster motivation and engagement to use the learning application. The perceived usefulness and intention to use provides promising results to use this tool as a feedback application in different learning settings.

**Argumentation Quality of Written Texts**

The mean number of arguments in texts from participants using ArgueLearn was 5.08 (SD= 1.76). From the text of participants using the alternative tool, I counted a mean of 3.2 arguments (SD= 1.51) (see Figure 43). A double-sided t-test confirmed that the treatment group wrote texts with a statistically significantly higher quality of formal argumentation: t-value= -3.622 and $p<0.001$. For the perceived quality of argumentation, I found that on a Likert scale from 1 to 5 points (1: very persuasive, 5: not persuasive) texts from the treatment group achieved an average value of 2.62 (SD= 0.96). Participants

using the alternative application wrote texts with a mean value of the perceived quality of argumentation of 3.21 (SD= 1.19). A double-sided t-test showed that the difference was statistically significant: t-value= -2.654 and p-value= 0.0105 ($p<0.05$). This clearly proves the hypothesis that individual feedback on students' argumentation helps them to write more convincing texts. The results show that students using ArgueLearn wrote texts with a better formal quality of argumentation as well as a better perceived quality of argumentation compared to the ones using the traditional approach.

#### 5.3.2.4 Qualitative User Feedback

As described above, I also included open questions in the survey to receive the participants' opinions about the tool they used. The general attitude for ArgueLearn was very positive. Especially the fast and direct feedback, the graph-like visualization of the argumentation structure, and the summarizing scores were mentioned quite often. However, sometimes ArgueLearn was not correctly classifying claims and premises, which users suggest improving. I translated the responses from German and categorized the most representative responses in Table 25.

#### 5.3.2.5 Discussion

The proof of concept evaluation demonstrated that adaptive and individual feedback on students' argumentation skills helps them to write more persuasive texts. Not only the perceived argumentation quality but also the formal quality of the argumentation was significantly higher for students using ArgueLearn compared to the ones using the alternative tool. Therefore, this study makes several contributions to current research. At the time of conducting the research, this study was one of the first to present evaluated design knowledge on how to build a learning tool to train argumentation skills based on adaptive and intelligent feedback. It provides a basis for researchers who also aim to develop learning tools to train metacognition skills to compare their solutions with ours. Lecturers and educational institutions can now use the design principles and findings to create their own learning tools for providing adaptive and intelligent support of argumentation skills in large-scale scenarios.

In the experiment, I prove the short-term influence of ArgueLearn on a student's argumentation skills. For the next evaluation, I suggest measuring the long-term learning effects on students' skills. This can be achieved with a longitudinal study in a real-world learning setting, e.g., in supporting the writing of peer reviews in business innovation

| Group | Feature |
|---|---|
| On user interaction | *"Easy handling. Especially marked texts with color and percentage values. Fast reaction of the tool. Motivated to write."* |
| On writing support | *"I was convinced by the mind map argumentation graphics. If you write longer texts, you can get lost quickly. With this tool, you can see how the sentences and argumentation stand together."* |
| On visualization | *"It was positive that the tool presented which premises support which of my claims and which arguments hung, so to speak, freely in the air."* |
| On graphics and color | *"I liked that the tool used colors to highlight the various elements and graphics to represent my text. Furthermore, the percentages of how my text was written were helpful and it is exciting to see how the tool judges my text."* |
| On the speed of the tool | *"Very fast and instant feedback."* |
| Improvements on feedback accuracy | *"I'm not sure how well this algorithm really understands what I'm writing."* |
| Improvements on user onboarding | *"Better introduction would be good, you had to try something before you knew how to do it. Maybe a short YouTube tutorial at the beginning, then you know more about what is important."* |
| Improvements on the summarizing scores | *"There are only three evaluation points (readability, coherence and persuasiveness) and it is unclear how relevant the individual factors are."* |

Table 25: Representative examples of qualitative user responses from users of ArgueLearn after study 1.
Source: Own illustration

lectures. Therefore, the next step will be to conduct a field experiment with three groups to evaluate the long-term impact of adaptive and intelligent feedback (provided by the feedback algorithm) on the development of students' argumentation quality (see study 2).

#### 5.3.2.6 Conclusion

In this research, I evaluated ArgueLearn, an AI-based argumentation writing support system, that provides students with feedback on the argumentation structure of a text by leveraging the recent advances of argumentation mining algorithms. I compared ArgueLearn to a proven argumentation writing support approach in a rigorous user study with 54 participants. I found that students using ArgueLearn wrote more convincing texts with a better formal quality of argumentation compared to the traditional approach. The perceived learning and intention to use provided promising results to use this tool as a feed-

back application in different learning settings. My results also offer design suggestions to further improve educational feedback applications based on intelligent algorithms. With NLP and ML becoming more powerful, I hope the work will attract other researchers to design and build more intelligent tutoring systems for other learning scenarios or metacognition skills.

### 5.3.3 Study 2: Proof of Value Evaluation of ArgueLearn in a Field Experiment

In study 1, I obtained positive results for a proof of concept that an AI-based argumentation writing support system based on my four design principles helps individuals improve their argumentation skills in short-term scenarios better than a proven scripting approach.

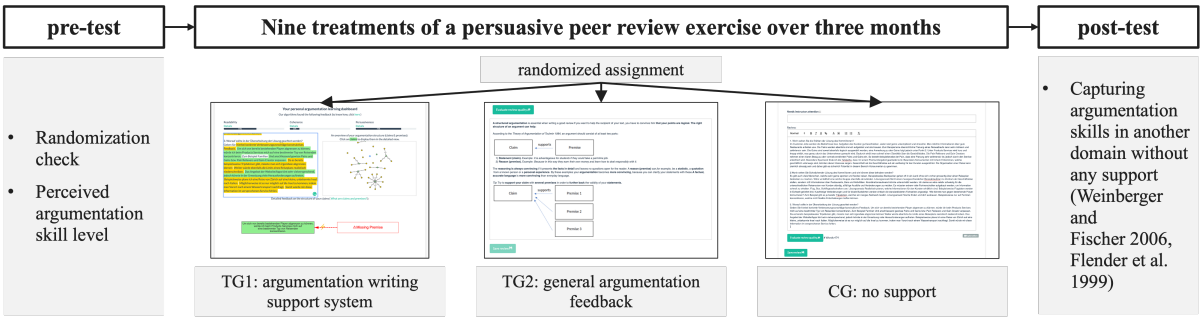



Figure 44: Overview of experimental procedure of study 2 to evaluate if an AI-based writing support system based on my principles, in comparison to static argumentation feedback, leads to better learning of argumentation.
Source: Own illustration

Hence, in study 2, I investigated further the extent to which an AI-based argumentation writing support system based on the design principles influences learners' long-term metacognitive argumentation skills in a large-scale learning setting. Therefore, I evaluated ArgueLearn with a proof of value evaluation in a real learning environment with 205 students in the context of higher education, from which 124 participated in the surveys. I implemented ArgueLearn in a business innovation lecture in a master's program at the University of St. Gallen in Switzerland. Following the Bauman and Tuzhilin (2018) design, I conducted a field experiment with three treatment groups to evaluate the impact of the AI-based writing support (provided by the back-end algorithm) on humans' argumentation quality. The functionalities necessary for the treatments were implemented in the learning system.

#### 5.3.3.1 Experimental Design

In the lecture in which I conducted the field experiment, students develop and present new business models. The students are asked to submit three assignments developing their own business idea over three months. Students are also required to peer review fellow students' assignments (three peer reviews per student per assignment). Since the course included three rounds of assignments, each student had to write nine peer reviews to elaborate on the strengths and weaknesses of their fellow students' business models and offer persuasive recommendations on what could be improved. The quality of the peer review (e.g., the persuasiveness) does not influence the students' final grades. However, submitting a peer review was mandatory for completing the assignments. For peer reviews, I provided the students with different levels of argumentative feedback on their persuasiveness. Figure 45 illustrates a screenshot of the task interface. The students were provided with a simple text editor where they were asked to enter their peer review. Below the text editor, a "evaluate review quality" button offered the students the option to receive feedback on their argumentation quality.

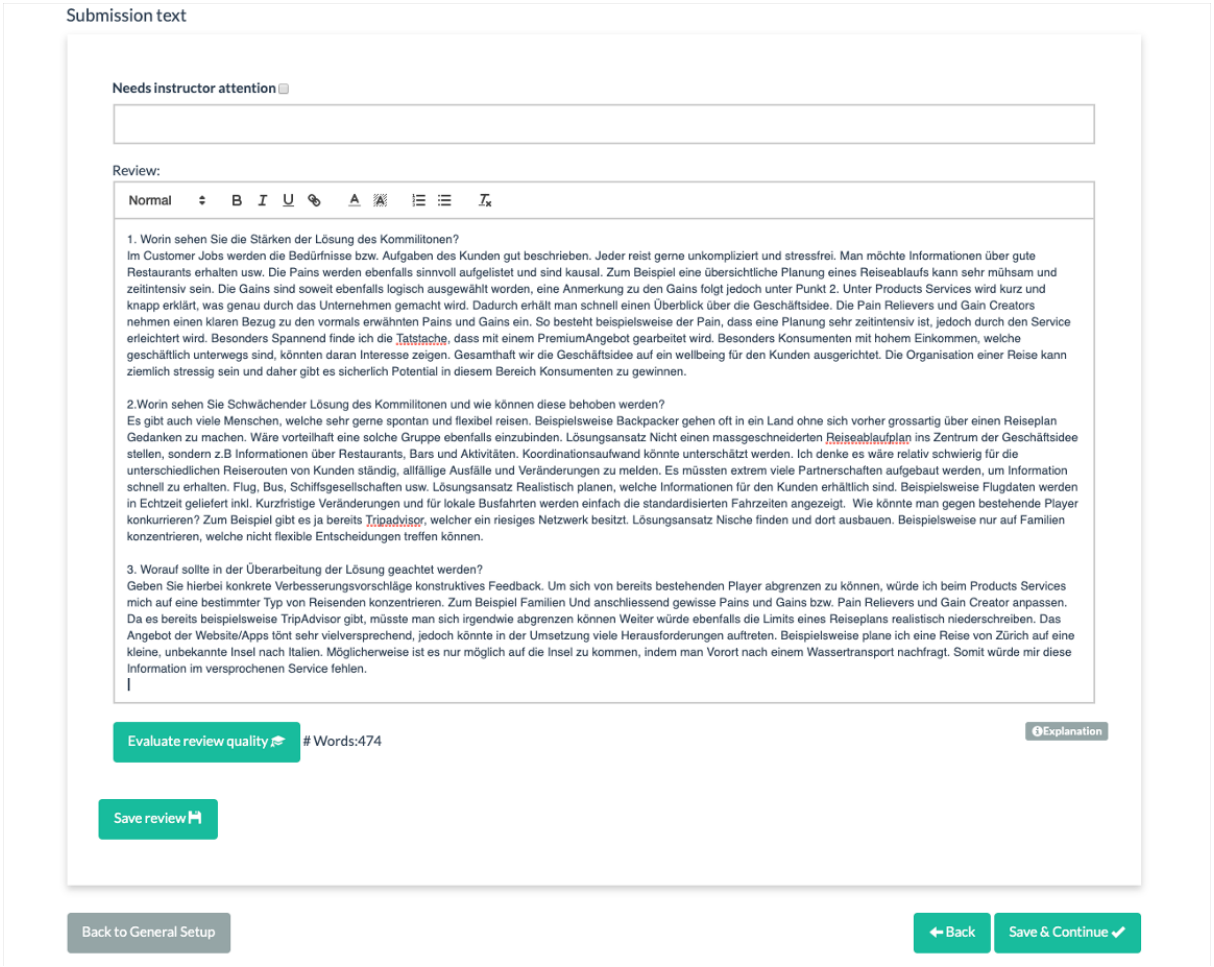



Figure 45: Screenshot of the task interface: students were asked to write a persuasive peer review.
Source: Own illustration

The field experiment comprised three main phases: (1) a pre-test phase, (2) an individual writing phase, and 3) a post-test phase. The pre-and post-test phases were consistent for all participants. In the writing phase, I manipulated the level of argumentation feedback participants received while writing their peer reviews.

**(1) Pre-test phase:** Before the assignment period, I conducted a mandatory post-survey to control for randomization. I checked the same items as in study 1 (i.e., personal innovativeness in the domain of information technology (Agarwal and Karahanna, 2000) and individuals' feedback-seeking (Ashford, 1986). Third, since no elaborated argumentation test was possible in the field setting, I asked the participants to determine their own perceived argumentation skill following the Toulmin model (Toulmin, 2003) and self-determination of competencies (Vansteenkiste and Ryan, 2013) using three items: "How do you rate your ability to argue?"; "How would you rate your ability to convince others?"; and "How would you rate your ability to write argumentative texts?". I wanted to control for argumentative competencies since I later measured the texts' formal argumentation quality. These items were measured using a 7-point Likert scale (1: extremely poor to 7: excellent; 4: fair).

**(2) Individual writing phase:** In the treatment phase, participants were asked to write nine peer reviews over three months. The first treatment group (TG1) received general (non-AI-based) argumentation feedback in the form of a short text and a visualization of a theoretical argumentation model (Toulmin (1984), see Figure 46); the second treatment group (TG2) received AI-based argumentation writing support by ArgueLearn following the design principles (see Figure 47). The control group (CG) received no feedback at all during the writing process. In total, 205 students were in the class and randomly assigned to one of the three groups. After three months, 124 students completed all three assignments. Table 26 presents an overview of the demographics.

**3) Post-test phase:** After the treatments, I asked the students to conduct a post-test via a web survey. The main objective was to measure the learners' metacognitive argumentation skills to evaluate the research question for study 2. Therefore, I tested whether the three groups' argumentation skills differ in a second argumentation domain. This is necessary to control whether the AI-based argumentation writing support system ArgueLearn significantly impacts the metacognitive argumentation skills or only humans' skills of writing more persuasive peer reviews (Toulmin, 2003). Thus, I asked the participants to read the same discussion as in study 1 between two teachers concerning the following topic: "Does TV make students aggressive?". I retrieved the topic with the discussion as well as the task from the construct of passive argumentative competency from Flender et al. (1999), since it is proven for measuring argumentative competencies. Similar to study 1, I asked the students to review both parties' argumentation (pro and contra) with

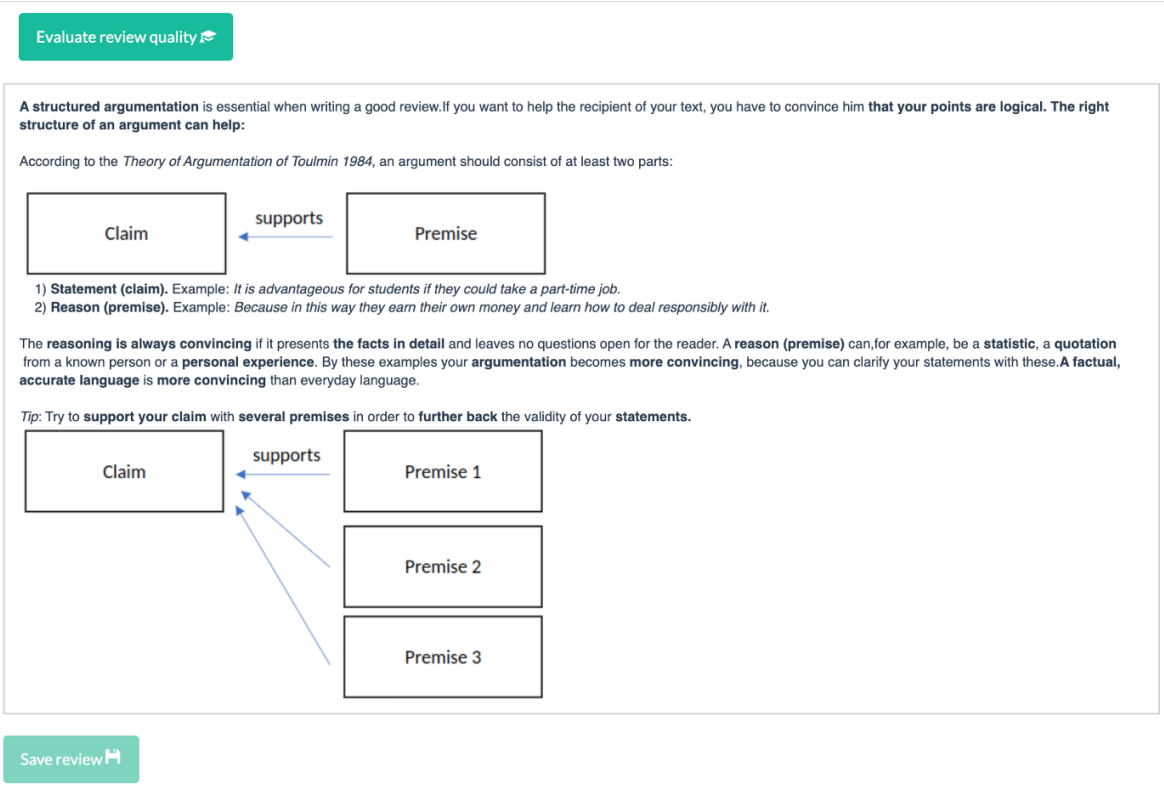



Figure 46: Screenshot of treatment group 1 (TG1): students received general argumentation feedback with hints and recommendations on what to improve based on the model of Toulmin (1984).
Source: Own illustration

respect to the weaknesses and strengths. The participants were told to spend at least 15 minutes writing this review. I captured the demographics and asked three qualitative questions: "What did you particularly like about the use of the argumentation tool?"; "What else could be improved?"; and "Do you have any other ideas?" to further evaluate the instantiations and to check for the completeness of the design principles.

#### 5.3.3.2 Results

Based on the written peer reviews from the post-test, I aimed to measure the students' formal quality of argumentation. I applied the same annotation scheme for argumentative knowledge construction used in study 1 (Weinberger and Fischer, 2006). Two annotators, who had already participated in the annotation process for the corpus, annotated the participants' texts based on the annotation guidelines and prior experience. As in study 1, I took the mean of the two annotators to judge the texts' formal and the perceived argumentation level.

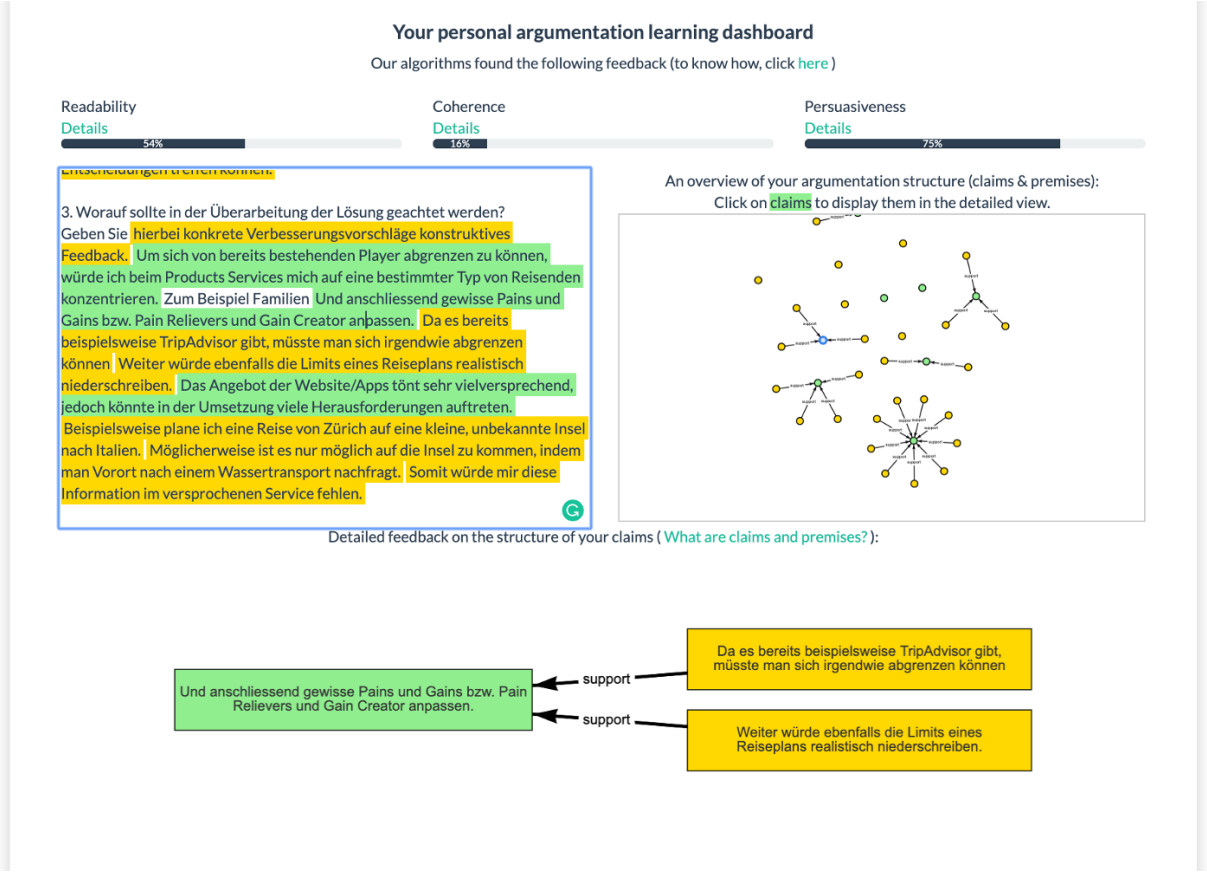



Figure 47: Screenshot of treatment group 2 (TG2): students received AI-based argumentation writing support by ArgueLearn following my four design principles.
Source: Own illustration

I compared the formal quality of argumentation between the written text of the post-survey (for an overview see Table 26 and Table 27). To clean the data, I deleted all answers of less than three sentences, since an argumentative discourse structure can only be reliably measured from a series of subsequent sentences. Therefore, from the 205 randomly assigned participants, I obtained 124 valid results. I used ANOVA analysis and checked their assumptions visually with a test for normality and a test for homoscedasticity, and all assumptions were met. To control for potential effects of interfering variables and to ensure successful randomization, I compared the differences in the means of the three constructs included in the pretest. For all three constructs, including personal innovativeness, feedback-seeking of individuals, and perceived argumentation skills, I received p-values larger than 0.05 between the three groups.

The number of formal arguments in texts from participants using ArgueLearn (TG2) is significantly higher than students receiving general feedback (TG1) and no feedback (CG). The ANOVA analysis confirmed that the TG2 participants wrote texts with a statistically significantly higher quality of formal argumentation: $p = 0.0009$ ($p < 0.001$). To measure

| Group | n | Formal argumentation | | TukeyHSD test TG1 and Cohen's d | TukeyHSD test (CG) and Cohen's d | Mean Age | Gender |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | | | | |
| **TG2: AI-based AWSS** | 41 | 4.04 | 1.33 | p = 0.015*<br>d = 0.6829 (medium) | p = 0.01**<br>d = 0.7395 (medium) | 25.40 | 21 female, 20 male |
| **TG1: General Feedback** | 36 | 3.2 | 1.09 | - | p = 0.8137<br>d = 0.1374 (negligible) | 25.70 | 8 female, 28 male |
| **CG: Control Group** | 47 | 3.03 | 1.40 | - | - | 25.27 | 18 female, 29 male |
| **Total** | 124 | p <0.001 | | ** p <0.01, * p <0.05 | | | |

Table 26: Mean and standard derivation of formal and perceived argumentation between the three groups of the field experiment (study 2).
Source: Own illustration

| Group | n | Perceived argumentation | | TukeyHSD test TG1 | TukeyHSD test (CG) |
|---|---|---|---|---|---|
| | | Mean | SD | | |
| **TG2: AI-based AWSS** | 41 | 3.06 | 0.81 | *p = 0.8486* | *p = 0.9886* |
| **TG1: General Feedback** | 36 | 2.73 | 0.91 | - | *p = 0.7842* |
| **CG: Control Group** | 47 | 3.07 | 0.99 | - | - |
| **Total** | 124 | *p = 0.8969* | | | |

Table 27: Mean and standard derivation of formal and perceived argumentation between the three groups of the field experiment (study 2).
Source: Own illustration

the effects between the groups, I run a Tukey Post-Hoc Test to check for differences in the formal quality of argumentation and confirmed a significant difference between participants using ArgueLearn (TG2) and participants receiving general argumentation feedback (TG1) with a p-value of 0.0150 (p >0.05). Moreover, the test also confirmed students in TG2 (ArgueLearn) wrote their text with better formal argumentation quality than the control group with a p-value of 0.001 (p <0.01). I also calculated the Cohen's d between the different groups to measure the effect size. Cohen suggested that d = 0.2 be considered a 'small' effect size while 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size. This means that if two groups' means do not differ by 0.2 standard deviations or more, the difference is trivial even if statistically significant (Cohen 1988). For the effect Between TG2 (ArgueLearn) and TG1, a d = 0.6829 indicates a medium effect size. The differences between TG2 and CG also indicate a medium effect (d = 0.739514) (Cohen 1988) for formal argumentation skills.

I found no significant difference in perceived argumentation quality between the groups ($p = 0.8969$, see Table 27). Participants using ArgueLearn (TG2) wrote their texts with perceived argumentation quality of 3.06 ($SD = 0.8077$), whereas TG1 students wrote with a perceived quality of 2.76 ($SD = 0.9140$). In the CG texts, I measured a perceived argumentation quality of 3.07 ($SD = 0.9998$). These results indicate that individual feedback on students' argumentation helps them to write more convincing formal texts in another domain (e.g., persuasive reviews about social debates). However, it does not help them to write with higher perceived argumentation levels in other domains. As described above, I also included open questions in the survey to obtain the participants' opinions on ArgueLearn's design and interactivity to further evaluate the functionalities and the design principles. The general attitude toward ArgueLearn was positive. Again, the fast and direct feedback mechanism (artifact 3, F3), the argumentation learning dashboard (artifact 4, F1-F6) and the graph-like visualization of the argumentation structure (artifact 4, F5) were praised several times. However, sometimes ArgueLearn did not correctly classify claims and premises, which users suggest needs improving (artifact 3, accuracy of the argumentation mining algorithm).

## 5.4 Discussion

The aim of study 2 was to evaluate the proof of value of the design instantiated in ArgueLearn. Based on a socio-technical perspective guided by self-regulated learning theory, I demonstrated how novel AI-based argumentation writing support systems can automate individual argumentation self-evaluation for individuals to learn and practice argumentation skills based on their logical argumentation errors irrespective of instructor, time, and location.

In this design science research project, I followed Peffers et al. (2007) to design, build, and evaluate ArgueLearn, an AI-based argumentation writing support system that provides students with adaptive feedback on the argumentation structure of a text independent of an instructor, time, and place by leveraging the recent advances of argumentation mining algorithms. I rigorously derived requirements from 138 scientific papers and from 30 semi-structured user interviews to formulate a concise set of design principles for AI-based argumentation tutoring systems. I evaluated those design principles as an instantiated artifact (ArgueLearn) with two empirical studies. I tested the impact of my design principles for AI-based argumentation writing support systems through ArgueLearn in a controlled laboratory experiment with 54 students (study 1) and in an extensive field experiment with 124 students over three months (study 2). I found that students using ArgueLearn improved their short- and long-term metacognition argumentation writing skills with a

better formal quality of argumentation compared to students who only received general or no writing support. The results indicate that students also write text with a higher perceived argumentation level in short-term scenarios. However, the perceived metacognitive argumentation level seems not to significantly improve through the AI-based argumentation tutoring system across different learning domains.

### 5.4.1 Theoretical and Practical Contributions

Besides the software artifact as a situated implementation of an AI-based argumentation tutoring system, I contribute design knowledge to the scientific knowledge base. Due to the systematic procedure, I aimed at generating a satisfying design contribution (Gregory and Muntermann, 2014). The resulting design knowledge is not only valid for the specific case but can also be transferred to further use cases in adaptive argumentation learning. For instance, it is easily possible to apply the concept of ArgueLearn in courses that deal with other content than business models or other languages. For this purpose, the back-end algorithm just needs to be adapted to the other scenario. As described, multiple corpora and argumentation mining models exist in literature that can be easily embedded in ArgueLearn, e.g., for English student essays (Stab and Gurevych, 2017) or English law cases (Mochales and Moens, 2008). The design principles of form and function and the overall system design do not need to be adapted for those use cases. Furthermore, it is also possible to transfer the design knowledge to pedagogical scenarios that target the training of other metacognition skills. For instance, if the learning of general feedback skills or empathy skills of students is aimed to be trained, a similar AI-based argumentation tutoring system can be used. However, in this case, the system design might need to be revised partially, e.g., the graph visualization or argumentation dashboard needs to be adapted. Due to this transferability of the design knowledge, the research does not only provide a Level 1 design science research contribution by showing a situated artifact implementation but also provides a nascent design theory (Level 2 contribution) (Gregor and Hevner, 2013). In addition to these contributions to the scientific knowledge base, the results are also valuable for the practical use of AI techniques such as NLP and ML-based feedback systems for education. the results indicate that the state of the art in NLP and ML techniques are suited to design complex adaptive skill learning systems that are able to support students individually – even in existing pedagogical scenarios such as peer reviews. I have thus shown that more advanced support possibilities in skill education can be implemented using techniques from NLP and ML in contrast to common e-learning approaches that are currently used.

### 5.4.2 Limitations and Future Work

However, the research also faces several limitations. For the aim of this study, I focused the research on students from the university. Even though it is reasonable to assume that the transferability to other cases is possible without major changes, I cannot prove it with the research design. Regarding the design, the main improvement potential lies in the accuracy of the feedback algorithm. I see two main options to tune the performance of the algorithm: (a) enrich the corpus with more annotated texts and (b) improve the performance of the models to enhance the performance through deep learning algorithms or transfer learning models like BERT (Devlin et al., 2018). Regarding, the implementation of the AI-based argumentation writing support system, I clearly do not want to replace human tutors, since I believe that skilled teachers will always be able to provide better adaptive writing support than ArgueLearn. However, I hope through the system human tutors can focus more on detailed questions and can devote more time to difficult cases. Moreover, further studies are necessary to reproduce my findings in this or in other domains. I also perceive several data privacy concerns regarding integration of our tool into a common writing editor, since personal data (e.g., about students' argumentation skills) might be exposed to third parties or systematically collected. Hence, I recommend future discussions on dealing with the trade-off of making novel user-centered learning applications widely accessible and easy to use without exposing learner data. Nevertheless, since our system is based entirely on native libraries and frameworks, I can ensure that our architecture ensures that no personal data leave the university server's infrastructure.

All in all, the research offers design knowledge to further improve educational feedback applications based on intelligent algorithms. With further advances in NLP and ML, I hope the work will attract researchers to design more intelligent tutoring systems for other learning scenarios or metacognition skills and thus contribute to the OECD Learning framework 2030 towards a metacognition-skill-based education.

## 5.5 Conclusion

In this section, I answered RQ1b, RQ2, and RQ3a stated in section 1. I started the research on the design of AI-based argumentation writing support systems by conducting a systematic literature review and 30 semi-structured interviews to systematically collect theory-based and user-centered design requirements for this class of argumentation learning systems. In this vein, I answered RQ1b. Next, I analyzed the collected design requirements and derived four design principles for the specific design of AI-based argumentation writing support systems according to the framework of Gregor et al. (2020)

and therewith answered RQ2.

Furthermore, I proposed a novel service-oriented architecture for AI-based argumentation writing support systems consisting of four artifacts. To demonstrate the feasibility of my design principles, I instantiated the proposed architecture by annotating a novel corpus of argumentative student-written peer reviews, by training an ML model to detect the annotated argumentation structures, and by developing a user-centered front end that provides students with adaptive argumentation writing support independent of an instructor, time, and location, leveraging the ML model. To answer RQ3a, I built ArgueLearn as an instantiation of a novel system class and conducted two empirical evaluation episodes following an ex-post summative evaluation strategy (Venable et al., 2016). I found that students using ArgueLearn according to my design principles improved their short- and long-term metacognition argumentation writing skills with a better formal quality of argumentation compared to students who only received general or no writing support.

Hence, this section forms one of the central contributions of my dissertation. The derived design principles, the corpus, and the design and architecture of ArgueLearn provide a foundation for the succeeding section on conversational argumentation learning (section 6) and on digital nudging for argumentation writing support on business pitches (section 7).

# 6 Designing Conversational Argumentation Learning Systems

In this section, I describe the design and evaluation of a pedagogical conversational agent for argumentation skill learning. Hence, I will directly answer RQ3b of this dissertation.

**RQ3b:** *What requirements should be considered when designing an adaptive conversational argumentation learning system that aims to support students in writing argumentative texts?*

**Method:** *Design science research including a series of experiments.*

**Results:** *Empirical evidence of the impact of conversational argumentation writing support system based on certain design principles on students' argumentation skills and their learning experience.*

I will investigate if an interactive pedagogical scenario (Chi and Wylie, 2014) based on a dialogue-based system (e.g., a conversational agent) helps students learn how to argue through individual tutoring, tasks scaffolding, and theory input. Based on the theoretical embedding of the ICAP-framework by Chi and Wylie (2014), my goal is to provide students with assistance comparable to a discussion with a human tutor. Therefore, I instantiate a conversational argumentation learning system with which I validate the hypothesis that an interactive dialogue-based system with adaptive argumentation feedback and tutoring helps student to learn how to argue (Chi and Wylie, 2014). I validate the artifacts and their influence on students' abilities to argue in two different experiments. To do so, I instantiate a conversational argumentation learning system called ArgueTutor according to theory-motivated and user-centered design principles.

Therefore, this section provides an insight into the design, development, and evaluation of the novel class of conversational argumentation learning systems. I found students using the instantiated conversational argumentation learning system ArgueTutor wrote more convincing texts with a better quality of argumentation compared to the ones using an alternative approach. The measured technology adoption constructs (such as level of enjoyment or ease of use) provide promising results to embed conversational argumentation learning systems in traditional learning settings. The results indicate that dialogue-based learning applications combined with NLP text feedback have a beneficial use to foster better writing skills of students.

In the following, I will outline the investigation of requirements for conversational argu-

mentation learning systems, describe the design and implementation of the conversational argumentation learning system ArgueTutor, and finally demonstrate and evaluate my design findings in two empirical studies. Afterwards, I will discuss the contribution and the limitations of this section. The resulting design principles and the empirical evaluation provide interesting findings and potential to leverage dialogue-based learning settings in different scenarios for metacognition skill learning.[11]

[11]The insights presented in this section are partly based on various papers (Wambsganss et al., 2021c, 2020f). I thank my collaborators, the reviews, and the Associate Editors for the valuable feedback on my work.

## 6.1 Introduction

One possible solution to imitate meaningful, individual instructor—learner interactions are pedagogical conversational agents (Winkler and Söllner, 2018) (see section 2 for theoretical background on pedagogical conversational agents). Due to developments in NLP and ML, pedagogical conversational agents are also becoming increasingly valuable for more creative and harder to grasp skills such as argumentation. By using an adaptive pedagogical conversational agent as a tutor for argumentation skill learning, based on recent developments in ML and NLP, students would be able to receive adaptive tutoring during the writing process autonomously and independently of the instructor, time, and place. However, current literature sparsely investigates an approach with principles and proof on how to design an adaptive dialogue-based learning system that guides and helps students to learn how to argue when writing a persuasive text.

In fact, at the time of conducting this research, argumentation mining has not been used for accessing the skill level of students to provide adaptive and individual tutoring through a dialogue-based interaction with a pedagogical conversational agent, such as providing adaptive feedback on the persuasiveness of texts or guiding a learner through the process of writing an argumentative text (Lawrence and Reed, 2019). Leveraging the technological advances of argumentation mining, NLP, and ML, I aim to address this gap by designing and evaluating a new form of human–computer interaction for argumentation skill learning. I aim to generate design knowledge for an adaptive dialogue-based argumentation learning tutoring system that assists students on an individualized level in writing argumentative texts and supports the development of their argumentation skills. My goal is to provide students with assistance comparable to a discussion with a human tutor.

Given this potential for leveraging a conversational learning tool in combination with argumentation mining, I followed the design science research methodology according to Hevner (2007) to design and build ArgueTutor (short for *Argumentation Tutor*), an adaptive dialogue-based tutoring system that provides students with adaptive and instant feedback, theoretical input, and step-by-step guidance during their writing process. Figure 48 shows the eight consecutive steps that have been conducted to ensure design knowledge is created using insights from the application domain and the knowledge base. I analyze requirements based on the environment (relevance) and knowledge base (rigor), derive design principles, and instantiate and evaluate them with my conversational argumentation learning system ArgueTutor in two consecutive studies (study 1 and study 2).

To collect requirements, I followed two different approaches: (1) a rigorous theory-motivated approach, where I systematically searched literature in the field of educational technol-

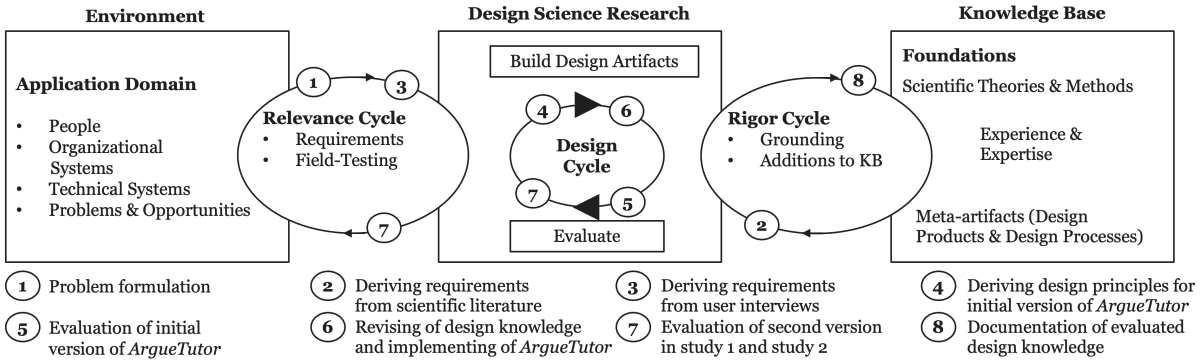



Figure 48: Overview of my methodology based on the design science research approach following Hevner (2007).
Source: Own illustration

ogy and human—computer interaction following (vom Brocke et al., 2015; Cooper, 1988) to carefully derive requirements and principles for a design of ArgueTutor, and (2), a user-centered design approach, where I conducted twelve semi-structured interviews with students to derive user stories. Based on these user needs, I built low-fidelity prototypes of ArgueTutor to test different design hypotheses with potential users to learn about the interaction flow of a dialogue-based learning tool for argumentation. With these two approaches, I present the final version of ArgueTutor.

To implement an adaptive dialogue-based learning tool, I (a) trained the intents of a conversational tutor based on rigorously scripted teacher–student conversations and (b) developed an argumentation mining model to assess the argumentation quality of student texts. To do so, I leveraged the argumentation annotated business-model peer review corpus from section 5, since the texts are derived from a pedagogical scenario and built on a rigorous annotation guideline with a moderate agreement. I trained and tuned a transfer learning model based on Devlin et al. (2018) to classify the argumentation quality of student texts. In combination with the trained chat intents, this model now serves as the underlying adaptive tutoring algorithm of ArgueTutor.

To evaluate ArgueTutor as an instantiated conversational argumentation learning system, I followed the evaluation framework proposed by Venable et al. (2016). In a pre-study, I performed a formative and artificial evaluation of the initial version to collect feedback in the early design stage of the dialogue-based tutoring system. Based on these findings, I refined the design knowledge, such as the instantiated design features, and built a fully functional software artifact as a second version of ArgueTutor, which supports and assists students on an individualized level in writing argumentative texts through guidance and adaptive argumentation feedback independent of an educator, time, and location.

Finally, I evaluate ArgueTutor following a naturalistic ex-post evaluation by Venable et al. (2016) in two empirical studies. In study 1, I compare ArgueTutor against a scripting approach for argumentation learning. I evaluated my conversational argumentation learning system in comparison with a discussion scripting approach that provided general argumentation feedback based on argumentation theory (Toulmin, 2003), a traditional approach for supporting persuasive writing in large-scale scenarios (e.g., Rietsche and Söllner (2019); Fischer et al. (2013); Scheuer et al. (2010)). In this study with 55 students, I observed that the participants who used ArgueTutor for a persuasive writing task wrote formally more argumentative texts. Furthermore, the perceived persuasiveness of these texts was significantly higher than of the texts from the traditional tool. I also measured the perceived ease of use and the level of enjoyment for both tools using key constructs (Venkatesh et al., 2003; Venkatesh and Bala, 2008). I found that both constructs provide promising results for the usage of ArgueTutor as a standard learning tool in lectures. The results suggest that ArgueTutor helps students to write more structured texts and motivates them to write more persuasive texts in peer learning settings, such as peer feedback scenarios. In study 2, I implement two different versions of ArgueTutor (one providing individual argumentation support, one providing non-adaptive argumentation support) in an experiment where 45 students had to write a persuasive text and received tutoring by ArgueTutor. The findings indicate that from a student's perspective the value of ArgueTutor not only lies in the dialogue-based interface and a step-by-step guidance but rather in the adaptivity of the pedagogical conversational agent's feedback

Hence, this study has three main contributions. First, with ArgueTutor I introduce the first instantiation of a prototype for the class of conversational argumentation learning systems. Second, I show the effectiveness of conversational argumentation learning systems through rigorously evaluating ArgueTutor in two empirical studies. The results provide insights into the benefits of leveraging NLP and ML for designing intelligent dialogue-based tutoring systems to foster argumentative writing in a student's learning journey. My results demonstrate an exemplary scenario of supporting metacognition skills in a scalable and individual way in possible large-scale or distance learning scenarios. Finally, I provide design knowledge for other researchers and educators to design and compare similar adaptive learning tools to foster the metacognition skill learning of students as a step to contribute to the OECD Learning framework 2030 towards a metacognition-skill-based education.

## 6.2 Theoretical Background

This study is based by previous research on technology-mediated learning systems for argumentation, by studies about argumentation mining algorithms, and by pedagogical conversational agents and the ICAP framework (Chi and Wylie, 2014), which serves as an underlying conceptual model for the main hypothesis (see section 2 for the theoretical background on these literature streams).

Conversational argumentation learning systems combine recent advances in NLP, ML and argumentation mining to build on new forms of human—computer interaction, such as adaptive pedagogical conversational agents, to intelligently tutor students in their individual argumentation learning process, e.g., with adaptive and instant feedback or theoretical input. Therefore, I build on adaptive support approaches to assess the potential of adaptive argumentation skill learning (Lawrence and Reed, 2019). In fact, the application of argumentation mining and adaptive dialogue-based tutoring systems has been motivated but rarely been investigated with design knowledge and empirical evaluation (Hobert and Wolff, 2019; Winkler and Söllner, 2018). I compare the novel system class against a nonadaptive static discussion scripting approach, since it is widespread and a common approach to foster students' argumentation skills in collaborative learning scenarios (e.g., Fischer et al. (2013); Rietsche and Söllner (2019)). Researchers have developed an increasing interest in intelligent writing assistance (Song et al., 2014; Stab and Gurevych, 2017), since it enables adaptive argumentative writing support with tailored feedback about arguments in texts (Wambsganss et al., 2020b; Lawrence and Reed, 2019). However, the complexity of using this technology for an adaptive dialogue-based learning system has been poorly assessed so far (Lawrence and Reed, 2019). In the approach, I focus on the first two subtasks to assess the argumentation level of a student to provide individual tutoring.

Following the underlying hypothesis of the ICAP framework by Chi and Wylie (2014) that interactive learning is always better that constructive, active, and passive learning, adaptive and interactive dialogue-based learning systems are capable of fostering the students' engagement as they add the new component of dialoging to technology-based argumentation learning systems. Compared to common argumentation learning systems, pedagogical conversational agents are able to discuss and tutor students through the learning content – just like human instructors would do. In fact, the successful application of pedagogical conversational agents to adaptively meet individual needs of learners and to increase their learning outcomes has been demonstrated for learning various skills, such as for problem-solving skills (Winkler et al., 2020), programming skills (Hobert, 2019), mathematical skills (Cai et al., 2019) as well as for learning factual knowledge

(Ruan et al., 2019), but it has not been investigated for argumentation skills. Therefore, I believe that a theory-motivated and user-centered design of an adaptive pedagogical conversational agent combined with intelligent algorithms to provide learning tutoring for argumentation skills by individually assisting students in writing persuasive texts would interactively foster learning according to the ICAP Framework.

## 6.3 Requirements for the Design of Conversational Argumentation Learning Systems

In this section, I will describe and discuss how I derive requirements for the system design of conversational argumentation learning systems. The problem formulation, described in the introduction and in the theoretical background, serves as the foundation for the derivation of the requirements.

### 6.3.1 Deriving Literature Requirements for Conversational Argumentation Learning Systems

To build a theory-motivated and user-centered learning tool, I followed two different approaches: a rigorous theory-driven approach and an agile user-centered approach following the build-measure-learn paradigm (Ries, 2011). For the rigorous theory-driven approach, I followed the approaches of Cooper (1988) and vom Brocke et al. (2015) to conduct a systematic literature review with the aim of deriving a set of theory requirements for the design of a dialogue-based argumentation learning system. I initially focused the research on studies that demonstrate the successful implementation of learning tools for argumentation skills and pedagogical conversational agents. The design of a conversational learning tool for argumentation skills is a complex project that is studied by psychologists, pedagogues and computer scientists with different methods. Therefore, I firstly concentrated on two main literature streams for deriving requirements: educational technology and human—computer interaction. I only included studies that dealt with or contribute to a kind of learning tool in the field of argumentation learning or dialogue-based learning systems, such as an established pedagogical theory. On this basis, I selected 85 papers for more intensive analysis. I have summarized similar topics of these contributions as literature issues and formed five clusters from them (e.g., LI1 (Soloway et al., 1994) or LI2 (Nass and Moon, 2000)). Based on these literature issues, I derived meta requirements for the design of a pedagogical conversational agent for argumentation learning. The resulting clusters are illustrated as literature issues in Figure 49.

### 6.3.2 Deriving User Requirements for Conversational Argumentation Learning Systems

Besides the rigorous theory-driven approach, I followed a user-centered design approach at the same time. As a start, I conducted twelve user interviews with students to receive an initial understanding of the needs and requirements of learners for a dialogue-based learning tool for argumentation. Therefore, I followed the expert interview method of Glaŝer and Laudel (2010). The interview guideline consists of 44 questions and each interview lasted around 30 to 45 minutes. The interviewees were students at the University of St. Gallen who are all potential users of an adaptive dialogue-based learning tool for metacognition skills. The participants were asked about the following topics: *experience with technology-based learning systems, perception of existing learning systems in use, experiences with conversational agents, requirements for an adaptive pedagogical conversational agent that supports writing argumentative texts* (e.g., functionalities, design), *requirements for an adaptive pedagogical conversational agent that supports learning how to argue through individual writing support* (e.g., functionalities, design).

The interviewed students were between 21 and 27 years old; nine were enrolled in a master's program and three in a bachelor's program at the university. All participants were business students; six were male, six were female. I recorded all interviews on audio and transcribed them. Based on the transcription I formatted abstract categories and identified 45 unique user stories in the twelve interviews. The coding was performed using open coding to form a uniform coding system during evaluation (Glaŝer and Laudel, 2010). Based on these results, I gathered user stories and aggregated the most common ones following Cohn (2004).

### 6.3.3 Deriving Design Principles for Conversational Argumentation Learning Systems

As illustrated, I have identified a concise set of literature issues, user stories, and formulated meta requirements and user requirements (see Figure 49). Based on these findings, I derived five design principles for adaptive conversational argumentation learning systems, a special class of adaptive dialogue-based learning tools for metacognition skills. For formulating the design principles, I relied on the conceptual schema of Gregor et al. (2020), and therefore believe they are self-explanatory. However, I will discuss them further in the evaluation. The design principles were followed in the instantiation of the current version of the conversational argumentation learning system ArgueTutor. The design principles included:

- For educational designers to design adaptive conversational agents for students to improve their argumentation independently of the instructor, time, and place, employ a PCA with guidance and support for argumentative writing embedded with social elements and an adaptive argumentative text feedback function to allow student to learn interactively.

- For educational designers to design adaptive conversational agents for students to improve their argumentation independently of the instructor, time, and place, employ a web-based conversational agent with a responsive, simple and functional UX to allow students to intuitively use the tool for learning tasks without any distraction.

- For educational designers to design adaptive conversational agents for students to improve their argumentation independently of the instructor, time, and place, employ a pro-active and individual argumentative tutoring and guidance with explanations based on an argumentation theory to allow students to receive support whenever they need.

- For educational designers to design adaptive conversational agents for students to improve their argumentation independently of the instructor, time, and place, employ an adaptive feedback function on argumentative texts with an analysis of individual argumentative components and feedback to allow students to assess their individual argumentation level.

- For educational designers to design adaptive conversational agents for students to improve their argumentation independently of the instructor, time, and place, employ a causal chat function with jokes or fun facts, to allow students to take a break from the primary learning activity.

## 6.4 Design and Implementation of a Conversational Argumentation Learning System

In this subsection, I will explain how I designed and implemented the two main components of a conversational argumentation learning system: the dialogue-based user interaction and the adaptive feedback algorithm in the back end. The basic user interaction concept of the instantiated conversational argumentation learning system ArgueTutor is illustrated in Figure 50. A user conducts a persuasive writing task and receives adaptive tutoring and feedback on the argumentation level.

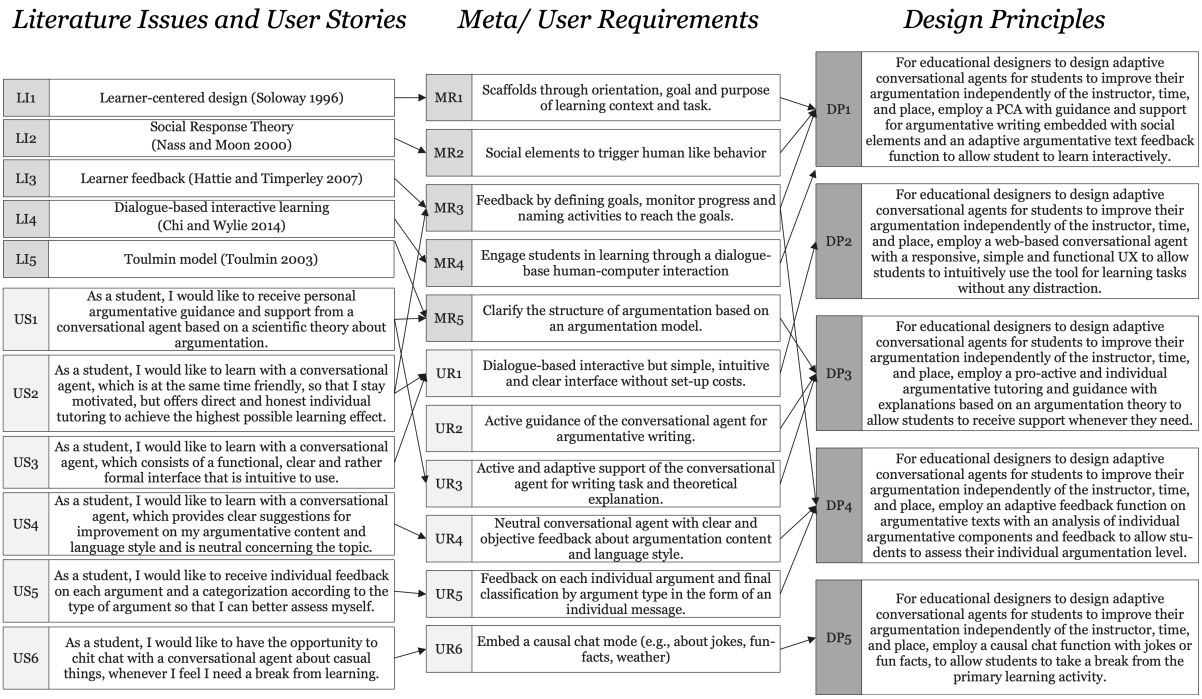



Figure 49: Overview of the derived design principles for conversational argumentation learning systems according to Gregor et al. (2020).
Source: Own illustration

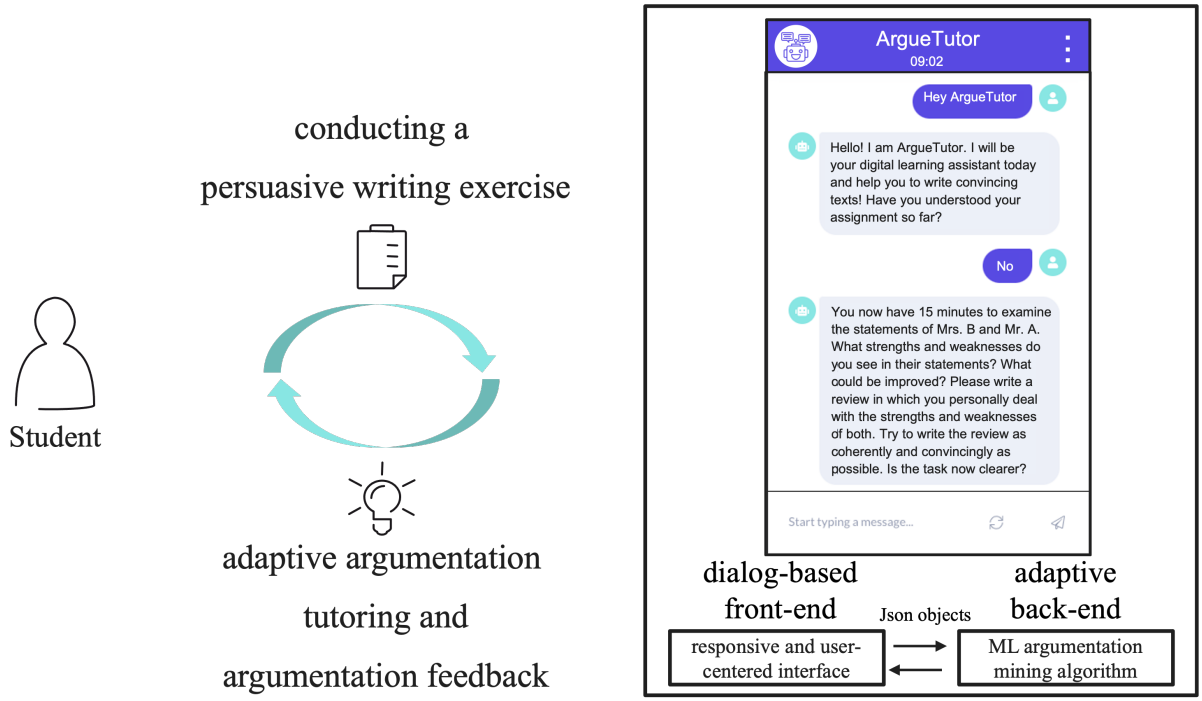



Figure 50: Basic user interaction concept of my conversational argumentation learning system ArgueTutor: a student is adaptively tutored through a writing exercise with individual argumentation feedback.
Source: Own illustration

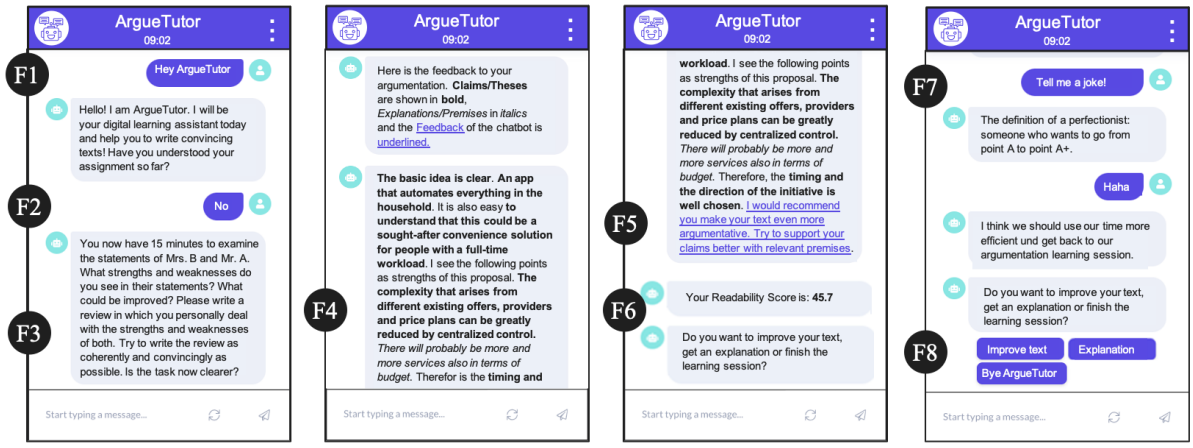



Figure 51: Different screenshots of the adaptive conversational argumentation learning system ArgueTutor: a user conducts a certain writing exercise and receives adaptive tutoring and feedback on the argumentation quality of her text.
Source: Own illustration

### 6.4.1 Instantiation of the Conversational Argumentation Learning System ArgueTutor

Following the design principles, ArgueTutor is built as a responsive web-based application that can be used on all kinds of devices. A screenshot of ArgueTutor and its core functionalities (e.g., *F1* - *F7*) can be seen in Figure 51 and in Figure 53. ArgueTutor was designed in German to provide German students with feedback on German texts. However, for ease of understanding the study, I translated parts of the user interface into English (e.g., see Figure 51). ArgueTutor consists of two modes: an adaptive learning mode and a casual chat mode. The dialog flow of ArgueTutor with the two interaction modes can be seen in Figure 54. The basic dialog flow of the adaptive learning mode of ArgueTutor was designed according to the didactical learning phases of *"Motivation, Difficulty, Solution, and Practice"* according to (Roth, 1970).

ArgueTutor guides students through a writing exercise with the aim to imitate a human educator (F1, F2, and F3). The pedagogical conversational agent proactively explains a writing task (e.g., students have to write persuasive peer feedback to a fellow student) (F3) and provides hints and explanations when the user asks for help such as argumentation theory input (Toulmin, 2003). Moreover, ArgueTutor is always able to provide individual feedback on the argumentation skill level of written student texts by highlighting argumentative components such as claims and premises (F4) and by tutoring students with an individual feedback message on what to improve depending on the student's skill level (F5). Inspired by Lippi and Torroni (2016b), claims are displayed in bold font, whereas premises are displayed in italic style (F4). Non-argumentative text paragraphs

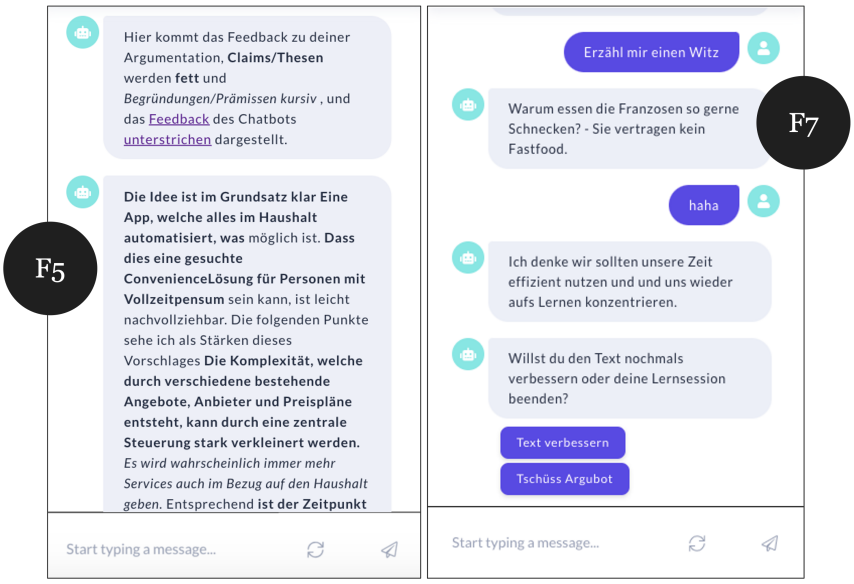



Figure 52: Screenshot of adaptive learning mode (F5) and casual chat mode (F7) of the conversational argumentation learning system ArgueTutor.
Source: Own illustration

are not highlighted. Additionally, ArgueTutor provides students with an individual summarizing feedback based on the number of premises and claims in the message (F5) (see Figure 52). For example, if the message contains less than two premises or contains more claims than premises, the user receives corresponding feedback indicating that the argumentation could be improved with certain improvement suggestions. Besides, based on user feedback, I embedded a readability score in the feedback to provide students with an overview of the quality of their general writing (F6). Therefore, I calculate the readability of the student's text based on the Flesch Reading Ease score (Flesch, 1943) and provide a small explanation for the individual score (e.g., if less than 40, the user receives a recommendations to improve the readability).

The interactions between the user and ArgueTutor are mixed between both typing and button selections. While writing a persuasive text for an exercise is typing based, selecting from multiple choices, asking for a task explanation or theory input are button-based. Predefined answer buttons help the user to receive an overview of the learning process and ensure both flexibility and efficiency regarding user interactions with ArgueTutor (F8). Moreover, ArgueTutor incorporates a user-initiated and rule-based casual chat mode. Students can ask ArgueTutor to tell jokes, fun facts, or talk about the weather to take a break from the primary learning activity (see Figure 52, F7) and thus change from the argumentation learning mode to the casual chat mode at any time they want. To imitate the students having a personal learning session with a human educator, I incorporated several more functions to incorporate real-world conversational elements into ArgueTutor's design. For example, I provided a wide variety of different responses

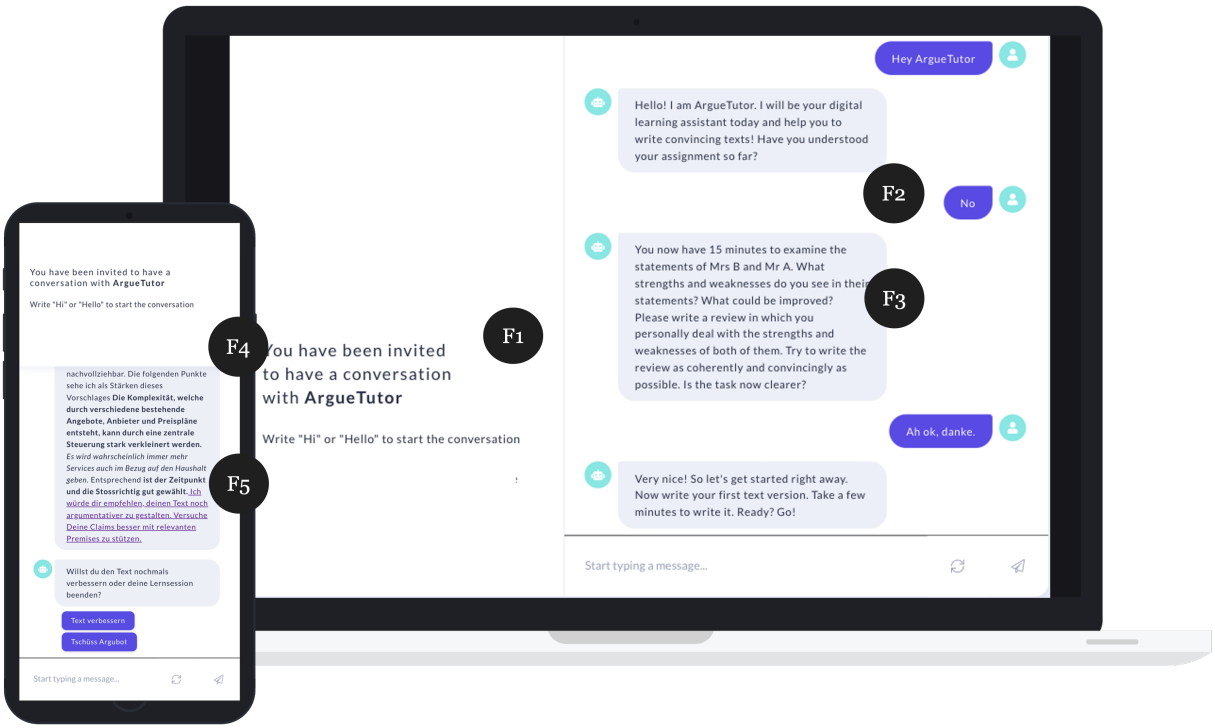



Figure 53: Screenshot of the interface of ArgueTutor with exemplars of design features in a mobile and desktop version.
Source: Own illustration

to common conversation states such as "how are you?" as well as positive reinforcement feedback that is typical of a study partner. The result of the dialog path of ArgueTutor from a user perspective is depicted in Figure 54.

The learning mode is based on a rule-based chat system combined with a supervised argumentation mining model to assess the argumentation level of students. The adaptivity of the argumentation feedback is implemented by training a model based on a corpus of persuasive student-written texts. The argumentation theory guidance, the task explanation and other tutoring functions are implemented through rule-based trained chat intents based on a word-to-vec model following the architecture of rasa nlu and rasa core (Bocklisch et al., 2017).

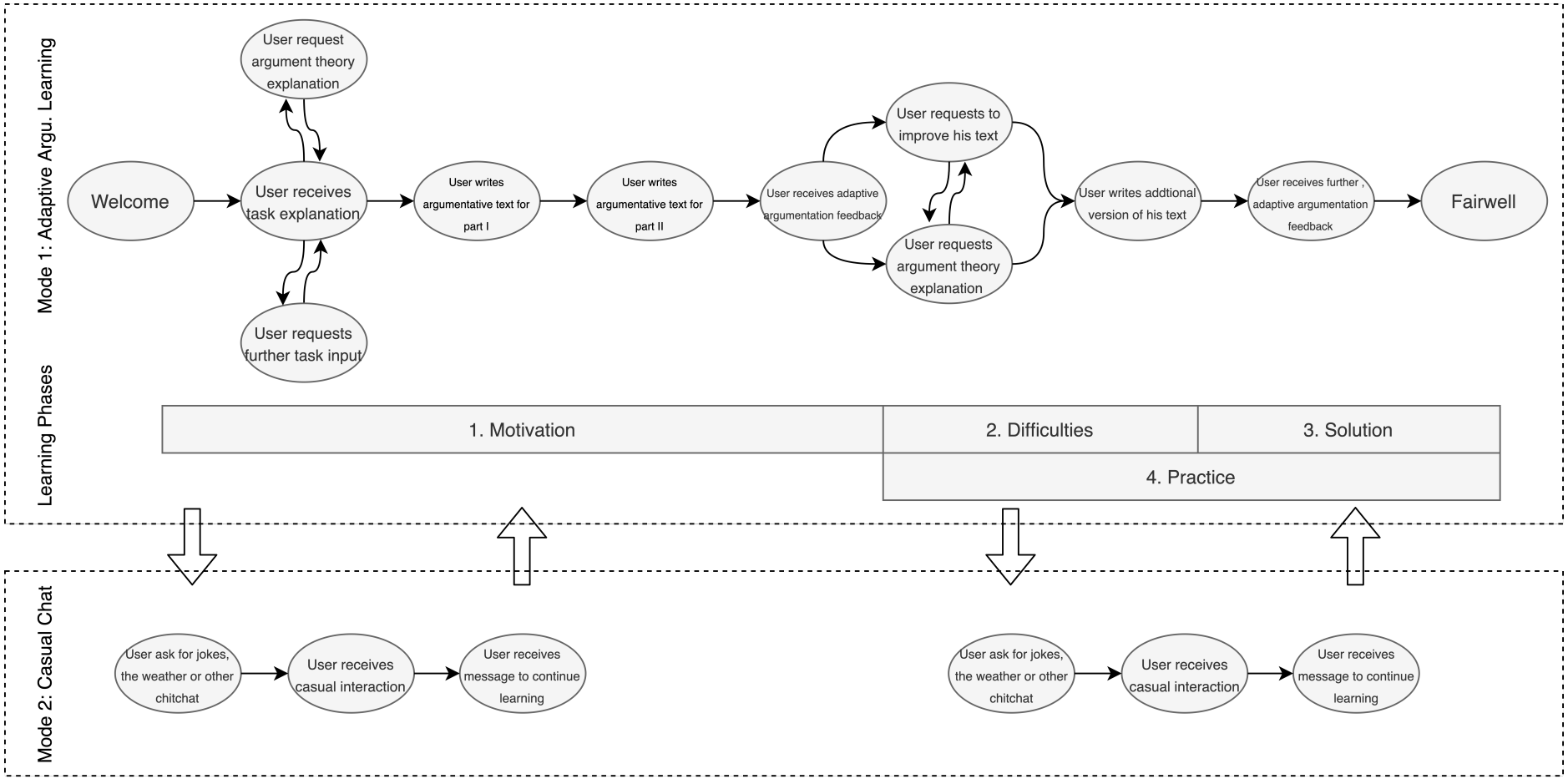



Figure 54: The dialog flow of ArgueTutor consists of an adaptive argumentation learning mode (mode 1) and a casual chat function (mode 2).
Source: Own illustration

All in all, I only included simple design elements in the design of ArgueTutor, since I received the user feedback to offer a more formal and functional tutoring experience that does not distract students. However, I believe these design principles might change depending on culture and context and can thus be easily adapted, e.g., by embedding ArgueTutor in a human persona with a name, face, picture, and the use of emojis.

### 6.4.2 Argumentation Mining Algorithm of ArgueTutor

To design an adaptive dialogue-based learning tool with individual and adaptive guidance, I trained and tuned a transfer learning model that fulfills the users' requirements to receive adaptive and instant tutoring and feedback on their texts.

A major prerequisite for developing supervised ML models based on NLP that are able to identify argumentative texts and argument components in written texts is the availability of annotated corpora. Hence, I leverage my German business model peer review corpus which I collected and annotated for the implementation of an AI-based argumentation writing support system (see section 5). The corpus consists of 1000 business model peer feedback essays written by students extracted from a large-scale lecture scenario. The texts are annotated for their argumentative components (claim, premise) as well as for the relations of the components.

Guided by recent literature about argumentation mining (Lawrence and Reed, 2019) and my findings about transfer learning for argumentation mining (see subsection 4.2), I trained a predictive model following the architecture of Bidirectional Encoder Representations from Transformers (BERT) proposed by Devlin et al. (2018). Hence, I classified text tokens as claim, premise, or as non-argumentative following the argumentation theory of Toulmin (2003). I used the BERT model from *deepset*[12], since it is available for German and provides a deep pretrained model that was unsupervised while training on domain-agnostic German corpora (e.g., the German Wikipedia). The novelty of this architecture is the ability to capture semantic information from pretrained texts, which can then be used for other downstream tasks without the need for retraining, e.g., for identifying argumentative components. For applying the model, the corpus texts were split into word tokens to fulfill the preparation requirements for BERT. The special preprocessing for BERT was conducted by utilizing the tokenizer and processor provided by deepset and by utilizing spacy[13]. The goal of the model is to provide accurate predictions to identify and classify argument components that can be used for accessing the skill level of students and thus provide adaptive guidance and feedback on how to improve their argumentation.

[12]https://github.com/deepset-ai/FARM
[13]https://spacy.io/

I split the data into 70% training, 20% validation, and 10% test data (Bird et al., 2009). For the proposed architecture, the input and output are adapted to the sequence classification task of argument component identification. The last hidden layer is a Recurrent Neural Network with 512 nodes that takes the BERT output and learns to feed into a sigmoid layer that classifies each token according to the predicted label. The proposed model was fine-tuned in several iterations and the best-performing set of hyperparameters included a learning rate of $5e^{-5}$, a warm-up and embedding dropout probability of 0.1 and 0.15 respectively. After several iterations, the final BERT model reached a macro f1 score of 73% for classifying text tokens into claim, premise, or non-argumentative tokens. An f1-score of 73% is a satisfying result when comparing to other studies on student-written argumentation identification. For example, Stab and Gurevych (2017) reached an f1-score of 73% for argumentation stance classification of student-written text in English. Moreover, I benchmarked the BERT model against bidirectional Long-Short-Term-Memory-Conditional-Random-Fields classifiers (BiLSTM-CRF), since several other authors reached satisfying results with a BI-LSTM-CRF for argumentation compound classification (e.g., Chernodub et al. (2019)). In combination with the corresponding embeddings vocabulary (GloVe) (Pennington et al., 2014) the LSTM only reached an unsatisfying f1 score of 57% (16% lower then the 73% of the BERT model).

## 6.5 Demonstration and Evaluation of a Conversational Argumentation Learning System

In this section, I demonstrate the application and evaluation of the design principles for conversational argumentation learning systems in the context of higher education. I provide an empirical evaluation of the proposed design principles in form of my instantiated conversational argumentation learning system ArgueTutor in two experimental studies (study 1 and study 2).

For the demonstration and the evaluation, I aim to formally validate the research artifacts' effectiveness in addressing the problem (Peffers et al., 2007; Venable et al., 2016). Hence, I performed a naturalistic ex-post evaluation of ArgueTutor in study 1 to evaluate if a conversational argumentation learning system helps students to improve their argumentation skills compared to a proven discussion scripting approach (Venable et al., 2016). In study 2, I implement two different versions of ArgueTutor (one providing individual argumentation support one, one providing non-adaptive argumentation support) in an experiment where 45 students had to write a persuasive text and received tutoring by ArgueTutor.

### 6.5.1 Pre-Study

Before conducting empirical evaluations with the final implementation of ArgueTutor, I designed low-fidelity prototypes of ArgueTutor to test different design hypotheses with end-users to learn more about the conversational flows and the human—computer interaction of an adaptive tutoring system for argumentation skills. Afterwards, I designed a clickable mock-up displaying a conversational interface with a step-by-step guidance for writing argumentative texts and individual argumentation feedback (without implementing trained chat intents in the back end). For the evaluation, I followed an ex-ante evaluation using an artificial evaluation setup as proposed by Venable et al. (2016). The purpose of the evaluation is to check whether the design principles are useful for learners, in order to incorporate any change requests. The design principles were specifically examined based on the criteria of usefulness and usability and evaluated by means of various questions.

I started the testing with seventeen low-fidelity paper prototypes and later two digital mock-ups of ArgueTutor. For example, I hypothesized that students aim to learn with a humanized conversational learning tool. I tested this hypothesis with three different paper prototypes. Prototype one was embedded in a rather functional design without incorporating humanized design elements such as a profile picture. Prototypes two and three were designed with a higher level of anthropomorphic elements (e.g., a profile picture and emoticons in a more colloquial conversation). The tutoring and feedback algorithm was simulated by a human. The hypothesis was validated with 16 users. However, I learned that the majority of students rather like a functional design with a low level of anthropomorphic elements. Therefore, the final version of ArgueTutor contributes to that with a rather functional design. Moreover, I hypothesized that students aim to chitchat with ArgueTutor during the learning process and prototype a chitchat function for ArgueTutor. I tested this with a low-fidelity chatbot using the platform WhatsApp and imitating a learning interaction (simulating the tutoring and feedback algorithm by a human). The hypothesis was validated with 16 users, which overall liked the idea of a chitchat function. However, I learned that students also want a tutor to actively motivate them to continue their learning process. Therefore, the final prototype ArgueTutor now contributes to that with two different chat modes (see Figure 54), a learning mode and a casual chit chat mode.

Afterwards, I created a questionnaire with 38 items with specific questions about the design principles, e.g., I asked questions about the perception of the argumentation theory input, the usability and usefulness of the adaptive tutoring functions, and the concept of the conversational argumentation learning system as a whole. The participants received a

link with a clickable mock-up of ArgueTutor as a web app. After testing the concept of my conversational argumentation learning system, the participants were asked to complete a survey. In total, 32 participants successfully tested my prototypical version of ArgueTutor and fully answered my survey. 22 participants were master students, eight were enrolled in a bachelor program, two had a different occupation. All were potential users of an adaptive dialogue-based argumentation tutoring system. Their age ranged from 22 and 34 years; 19 were female, 13 male.

In order to obtain detailed feedback on the design principles, the respondents were able to classify the usefulness of the design principles based on a 7-point Likert Scale, e.g., for usability, ranging from "completely useful (1)" to "completely useless (7)". The evaluation of the general usability, as well as usefulness, yielded the following result: The respondents rated the implemented design and the design principles positively compared to the neutral value of 4 (mean of 2.45 of all design principles related questions). The overall concept of an adaptive pedagogical conversational agent for argumentation learning was also rated as very useful (mean of 2.16, n=32). For example, the argumentative guidance and theoretical argumentation input were rated by 93.5% as very useful. Moreover, I asked the participants for some qualitative feedback and additional ideas to improve the concept of ArgueTutor. Almost all users mentioned that the accuracy of the adaptive argumentation feedback is very important for revising one's own argumentation quality of a text and thus increasing their argumentation.

In total, I conducted three cycles with a total of 77 different users (16 in cycle one, 16 in cycle two, and 45 in cycle three) for the pre-studies. These users were different to the ones recruited for the semi-structured interviews but also students from the university with a similar age and gender distribution.

### 6.5.2 Study 1: Proof of Concept Evaluation of ArgueTutor

In this subsection, I describe the experimental setup for study 1. The goal was to evaluate the hypothesis that adaptive tutoring on students' argumentation will help them to write more convincing texts. To evaluate the hypothesis, I designed an experiment in which participants were asked to write a peer review based on a given essay. Participants were randomly assigned to a treatment and a control group. The treatment group used ArgueTutor to do the writing exercise, while participants in the control group used the alternative learning tool. I recruited 55 students from the University of St. Gallen to take part in the experiment. The experiment was conducted as a web experiment facilitated by the behavioral lab of the university. After randomization, I counted 31 valid results in the treatment and 24 in the control group. Participants of the treatment group had an average

age of 22.75 (SD= 1.96), 17 were male, 14 female. In the control group, participants' average age was 23.52 (SD= 2.81), 11 were male, 13 were female. All participants were compensated with an equivalent of about 12 USD for a 25- to 30-minute experiment.

#### 6.5.2.1 Alternative Argumentation Learning Tool

To evaluate ArgueTutor, I compared it with a nonadaptive discussion scripting application, since this approach was empirically proven to foster the formal quality of the argumentation of students (Stegmann et al., 2012; Scheuer et al., 2010). To control the differences and similarities in the design between the alternative tool and ArgueTutor, I also built the discussion scripting approach myself (see Figure 55). The learning tool supports the writing process of users with theory-based argumentation input and general nonadaptive theory-based recommendations on how to improve the argumentation level. Users can access the theory input and the general writing recommendations by clicking the "here" button. The theory input and argumentation writing recommendations are then displayed next to the writing editor. To ensure that ArgueTutor and the argumentation writing system are consistent with each other, there are many functionalities that are shared between them. First, the introduction text is the same across both apps. Moreover, the theory input and the general argumentation recommendations are the same for both learning tools following the Toulmin model (Toulmin, 2003).

#### 6.5.2.2 Experimental Design

The experiment consisted of three main parts: (1) pretest, (2) writing exercise and (3) posttest. The pre- and post-phases were consistent for all participants. In the writing phase, the treatment group used ArgueTutor to conduct a persuasive writing exercise, whereas participants of the control group conducted the same exercise using the alternative tool.

**(1) Pretest:** The experiment started with a pre-survey of 14 questions. Here, I tested three different constructs to assess whether the randomization was successful. First, I asked four items to test the personal innovativeness in the domain of information technology of the participants following (Agarwal and Karahanna, 2000). Exemplary items were *"I like to experiment with new information technologies"* or *"If I heard about a new information technology, I would look for ways to experiment with it,"*, *"In general, I am hesitant to try out new information technologies"* and *"Among my peers, I am usually the first to try out new information technologies".* Second, I tested the construct of feedback-

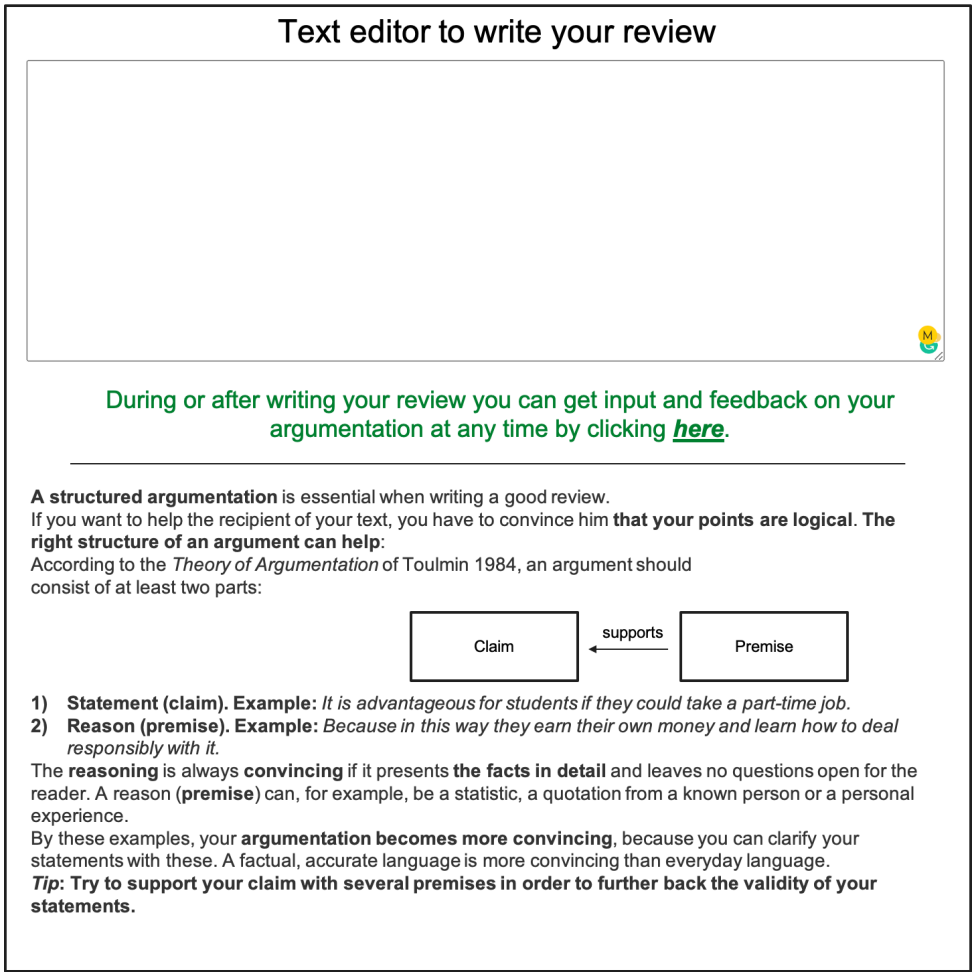



Figure 55: Basic user interaction concept of alternative argumentation learning tool: students receive theory input and general recommendations on the argumentation of a given text.
Source: Own illustration

seeking of individuals following (Ashford, 1986). Example items are: "*It is important for me to receive feedback on my performance.*" or "*I find feedback on my performance useful.*" Both constructs were measured with a 1- to 5-point Likert scale (1: totally disagree to 5: totally agree, with 3 being a neutral statement).

Similar to my experiments in section 5, I controlled for the argumentative competencies of the participants, since I later measured the formal and perceived quality of the argumentation of the written texts. Therefore, I captured the construct of passive argumentative competency following (Flender et al., 1999), as it is a proven construct to measure argumentative competencies in German. Participants were asked to read a discussion between two teachers concerning the topic "*Does TV make students aggressive?*". I retrieved the entire discussion with the topic and the measurements from (Flender et al., 1999). Based on the discussion, I asked the participants three questions concerning the

argumentation structure and the content of the text with multiple choice answers: "*What kind of argumentation style or structure is used?*", "*How can a new argument be added to the discussion?*" and "*Which of the following standpoints do both parties agree on?*" (Flender et al., 1999). Additionally, participants were asked how sure they were about the answers on a 1- to 5-point Likert scale (1: very sure, 5: not very sure, with 3 being a neutral statement). The competencies were then assessed by calculating a score from 0 to 27 following the measurements of (Flender et al., 1999).

**(2) Writing exercise:** In the writing part of the experiments, I asked the participants to conduct a persuasive writing task, simulating a typical student essay homework. Similar to study 1 in section 5, I asked the students to write a review about the discussion from the pre-survey. Therefore, students were asked to assess the argumentation of both parties (pro and contra) concerning the weaknesses and strengths of their argumentation. The participants were told to spend at least 15 minutes writing this review. A countdown indicated them the remaining time. They were only able to continue the experiment after the countdown was finished. The treatment group was using ArgueTutor to write the review, the control group was using the reference tool. I did not provide any introduction to any of the tools. The students using ArgueTutor were adaptively tutored through the writing exercise, e.g., through theory input, individual recommendations, and adaptive argumentation feedback based on the feedback algorithm. Participants using the reference tool retrieved help based on argumentation theory input and general argumentation recommendations during the writing process following (Toulmin, 2003).

**(3) Posttest:** In the post-survey, I measured the perceived level of enjoyment of the students, since enjoyment during a learning process has a major influence on the adoption of IT tools (Lee et al., 2005) and on the learning success of students (Pekrun and Stephens, 2012). Therefore, I asked the students the following items: *"The interaction with the learning tool was exciting"* and *"It is fun to interact with the learning tool"* (Kim et al., 2019). Moreover, I measured the perceived ease of use of the participants following the technology acceptance model of (Venkatesh et al., 2003; Venkatesh and Bala, 2008) and captured the demographics. In total, I asked 13 questions. Example items for the three constructs are: "*Imagine the tool would be available in your next course, would you use it?*", "*The use of the argumentation tool enables me to write better argumentative texts.*" or "*I would find the tool to be flexible to interact with*". Finally, I asked three qualitative questions: "*What did you particularly like about the use of the argumentation tool?*", "*What else could be improved?*" and "*Do you have any other ideas?*"

### Measurement of Argumentation Quality

Besides measuring the ease of use and the level of enjoyment, the main objective was to

measure the quality of the written texts from both groups to evaluate the main hypothesis. Therefore, I measured the formal quality of argumentation and the perceived quality of argumentation as done already for the evaluations of AI-based argumentation writing support systems in section 5.

**(1) Formal quality of argumentation**: The written peer reviews were analyzed for the formal quality of argumentation. I applied the annotation scheme for argumentative knowledge construction described by Weinberger and Fischer (2006). This annotation scheme was applied in various studies and has proven high objectivity, reliability, and validity (e.g., Stegmann et al. (2012)). To measure the formal quality of argumentation, the annotator had to distinguish between (a) *unsupported claims*, (b) *supported claims*, (c) *limited claims,* and (d) *supported and limited claims*. A more precise description of the scheme can be found in Weinberger and Fischer (2006). Three annotators were trained based on my 15-page annotation guideline of the corpus from section 5 to assess the argumentation components of persuasive reviews. The formal quality of argumentation of the individual user was then defined by the number of arguments written by a user during the writing phase. Following Stegmann et al. (2012), only *supported*, *limited* and *supported and limited claims* were counted as argumentation.

**(2) Perceived quality of argumentation:** The perceived quality of argumentation was annotated by the same three annotators. The objective was to subjectively judge how persuasive the given argumentation is on a Likert scale from 1 to 5 points (1: not very persuasive, 5: very persuasive). I took the mean of all three annotators as a final variable for the formal and the perceived quality of argumentation of the texts.

#### 6.5.2.3 Results

To evaluate the hypothesis that adaptive tutoring on students' argumentation will help them to write more convincing texts, the objective was to answer two research questions (RQ):

**RQ1:** *How effective is ArgueTutor at helping users to write more persuasive texts compared to the traditional approach?*

**RQ2:** *Do students perceive the interaction with ArgueTutor as easy and enjoyable to use during their writing process, and would they continue to use it in the future?*

To evaluate the first research question, I compared the formal quality of argumentation and perceived quality of argumentation between the written text of the treatment and

| Group | Formal argumentation | Perceived argu. |
|---|---|---|
| **Mean ArgueTutor** | 3.56 | 3.48 |
| **Mean reference tool** | 2.64 | 2.80 |
| **SD ArgueTutor** | 1.81 | 0.83 |
| **SD reference tool** | 1.21 | 1.43 |
| **p value** | 0.03459 | 0.03961 |

Table 28: Results of formal and perceived quality of argumentation between both tools. Source: Own illustration

the control group. Therefore, I applied a *Welch Two Sample t-test* to evaluate whether the means of the constructs are significantly different between the groups.

The second research question was answered by comparing the constructs of perceived ease of use and level of enjoyment for participants using ArgueTutor compared to participants using the alternative tool. I performed a *Welch Two Sample t-test* to assess whether differences between both groups are statistically significant. Moreover, I compared the results of ArgueTutor to the midpoints scale to validate a general positive technology acceptance. To ensure that the randomization resulted in randomized groups and to control for potential effects of interfering variables with the small sample size, I compared the differences in the means of the three constructs included in the pretest. For all three constructs, including personal innovativeness, feedback-seeking of individuals and passive argumentative competency, I received p values larger than 0.05 between the treatment and the control groups (for personal innovativeness p= 0.1436, for feedback-seeking of individuals p= 0.7537 and for passive argumentative competency p= 0.8495). This demonstrated that no significant difference in the mean values for these three constructs exists between the groups.

**Argumentation Quality of Written Texts**

The mean number of arguments in the texts from participants using ArgueTutor for the writing exercise was 3.56 (SD= 1.81). For the texts from participants using the alternative static tool, I counted a mean of 2.64 arguments (SD= 1.21) (see Table 28). A double-sided t-test confirmed that the treatment group wrote texts with a statistically significantly higher quality of formal argumentation: t value= 2.1738 and p value= 0.03459 ($p<0.05$). For the perceived quality of argumentation, I found that on a Likert scale from 1 to 5 points (1: not very persuasive, 5: very persuasive) texts from the treatment group achieved an average value of 3.48 (SD= 0.83). Participants using the alternative application wrote texts with a mean value of the perceived quality of argumentation of 2.80 (SD= 1.43). A double-sided t-test showed that the difference was statistically significant: t value= 2.114 and p value= 0.03961 ($p<0.05$). This clearly proves the hypothesis that adaptive dialogue-based tutoring during students' argumentative writing process helps them to

write more convincing texts. The results show that students using ArgueTutor wrote texts with a better formal quality of argumentation and with a better perceived quality of argumentation compared to the ones using the static traditional approach.

**Students' Perception of Ease of Use and Enjoyment**

| Group | ease of use | level of enjoyment |
|---|---|---|
| **Mean ArgueTutor** | 3.73 | 3.41 |
| **Mean reference tool** | 3.45 | 3.00 |
| **SD ArgueTutor** | 0.64 | 0.89 |
| **SD reference tool** | 0.69 | 0.88 |
| **p value** | 0.0286 | 0.02135 |

Table 29: Results of the perceived ease of use and level of enjoyment ArgueTutor and the reference tool on a 1 - 5 Likert Scale.
Source: Own illustration

To evaluate the students' perception, I calculated the means of the perceived ease of use and the level of enjoyment. I compared the results of ArgueTutor with the results of the alternative tool. The perceived ease of use of ArgueTutor was rated with a mean value of 3.73 (SD= 0.64) and the average perceived level of enjoyment of ArgueTutor was 3.41 (SD= 0.89). These values are significantly better than the results of the alternative approach. For perceived ease of use I observed a mean value of 3.45 (SD= 0.69) and for perceived level of enjoyment, the value was 3.00 (SD= 0.88) for participants from the control group. The results show that the participants of the experiment rated the ease of use of ArgueTutor as an adaptive dialogue-based tutoring system positively compared to the usage of the alternative application. The statistical significance was also proven in a double-sided t-test for all three constructs (see Table 29). Moreover, the mean values of ArgueTutor are also very promising when comparing the results to the midpoints. All results are better than the neutral value of 3, indicating a very positive value for the level of enjoyment and perceived ease of use. A high level of enjoyment and ease of use is especially important for learning tools to ensure students are experiencing joy in the usage of the tool and they find it easy to interact with. This will foster motivation, engagement, and adoption to use the learning application.

#### 6.5.2.4 Qualitative User Feedback

I also asked open questions in the survey to receive the participants' opinions about the tool they used. The general attitude for ArgueTutor was quite positive. Participants positively mentioned the fast and adaptive feedback (F4 and F5), the simple conversational interaction flow, and the adaptive feedback message with the readability score (F5 and

F6) several times. However, participants also asked if ArgueTutor could provide even more detailed feedback about the argumentation (e.g., through displaying argumentative relations between the components) and provide transparent explanations on how the feedback mechanism works. I translated the responses from German and clustered the most representative responses in Table 30.

| **Cluster** | **Feature** |
|---|---|
| On the user interaction | *"I liked the interactive dialogue-based interaction. It made the task more interesting compared to traditional writing tools with feedback. In a way, you don't feel alone when writing your essay and you get a second try for improvements."* |
| On feedback accuracy | *"The feedback was pretty accurate and criticized the points I would have criticized myself."* |
| On readability score and feedback message | *"The readability score and the recommendation seems to be correct and helped me to reflect on my argumentation."* |
| On the feedback visualization | *"I found it very revealing that different fonts were used to show how my argumentation is structured. This function creates added value."* |
| On speed of the tool | *"Very fast reaction time of ArgueTutor."* |
| Improvements on transparency | *"It was not quite clear to me how the feedback algorithm worked and what the score exactly measures."* |
| Improvements on feedback granularity | *"The feedback regarding my text could be more precise and detailed."* |

Table 30: Representative examples of qualitative user responses.
Source: Own illustration

#### 6.5.2.5 Discussion

This research study illustrated that adaptive dialogue-based tutoring on students' argumentation skills during a writing exercise helps them to write more persuasive texts. The perceived and the formal argumentation quality was significantly higher for students using the instantiated conversational argumentation learning system ArgueTutor compared to the ones using an alternative discussion scripting approach for the exact same writing exercise. I believe that the ICAP framework could explain the results. Accordingly, a dialogue-based interactive learning journey increases the engagement of the students compared to an active learning scenario. Therefore, the user-centered and theory-based design of an adaptive conversational argumentation learning system combined with intelligent algorithms interactively fosters learning according to the conceptual model of the ICAP framework (Chi and Wylie, 2014). Also, the perceived ease of use and the level of enjoyment showed positive results for the usage of a learning tool in a real-world scenario. Especially a high level of enjoyment during the learning process is important

for the long-term adaption of such learning tools since this is proven to foster motivation and engagement in the learning process.

#### 6.5.2.6 Conclusion

In the research study, I compared a novel conversational argumentation learning system, ArgueTutor, to a discussion scripting approach in an experiment with 55 participants. I found that students using ArgueTutor to conduct a writing exercise wrote more convincing texts with a better formal quality of argumentation compared to the traditional approach. The perceived ease of use and enjoyment offers promising results to use ArgueTutor as a learning tool in different learning scenarios. All in all, the research offers design knowledge to further improve dialogue-based tutoring systems based on techniques from NLP and ML.

In the next subsection, I will build on my findings from study 1 and evaluate two different designs of ArgueTutor to investigate the specific effect of adaptive feedback in the dialogue-based interaction of conversational argumentation learning system and its influence on user perception.

### 6.5.3 Study 2: Experimental Evaluation of Adaptive Conversational Feedback

After evaluating the general effectiveness of conversational argumentation learning systems in comparison to traditional static argumentation learning approaches, the objective of study 2 is to dive deeper into the adaptive interaction of conversational argumentation learning systems. My goal is to evaluate the effectiveness and usefulness of the adaptive feedback of a conversational argumentation learning system with the instantiation of ArgueTutor. To do so, I designed a similar experiment as done in study 1, in which participants were asked to write a persuasive peer review based on a given essay. Participants were randomly assigned to two different groups in which I manipulated the level of adaptivity of my conversational argumentation learning systems. The treatment group (TG) used the fully functional version of ArgueTutor (with adaptive argumentation feedback on texts, including F5 and F6), while participants in the control group (CG) used a version of ArgueTutor that only provides general feedback on texts (without my trained argumentation mining feedback intend – not including F5 and F6). Instead, participants of the CG received random general argumentation feedback based on the argumentation model of Toulmin (2003), e.g., “please make sure all claims are sufficiently supported by

at least one premise".

In total, I recruited 45 students from the University of St. Gallen through social networks and mailing lists to take part in this experiment. After randomization, I had 22 participants in the treatment and 23 in the control group. The participants received a link to the experiment, which was conducted in the browser using the tool unipark. Participants of the treatment group had an average age of 24.85 (SD= 2.32), 17 were male, 5 were female. In the control group, participants' average age was 25.39 (SD= 0.98), 14 were male, 9 were female. On average the experiment took 25 to 35 minutes.

#### 6.5.3.1 Experimental Design

Similar to study 1, the experimental design consisted of three main phases: (1) pretest phase, (2) individual writing phase and (3) post-test phase. The pre-and post-phases were consistent for all participants. In the writing phase, I manipulated the level of argumentation feedback participants received during writing their peer reviews.

**Pretest Phase:** As in study 1, the experiment started with a pre-survey with 14 questions. I checked the same items as in study 1 including the three different constructs of personal innovativeness in the domain of information technology (Agarwal and Karahanna, 2000), feedback-seeking of individuals (Ashford, 1986), and passive argumentative competency (Flender et al., 1999) to assess whether the randomization resulted in randomized groups (exemplary items are described in study 1 of this section).

**Individual Writing Phase:** In the writing phase of the experiments, I asked the participants to write a review about the argumentation of both parties (pro and contra) concerning the weaknesses and strengths of their argumentation, similar as done in study 1. The participants were told to spend at least 15 minutes writing this review. A countdown indicated the remaining time. They were only able to continue the experiment after the countdown was finished. The treatment group used ArgueTutor with adaptive argumentation tutoring to write the review, the control group used the version of ArgueTutor without the adaptive argumentation tutoring, only receiving general nonadaptive support based on the Toulmin theory (Toulmin, 1984). I did not provide any introduction to any of the tools.

**Posttest Phase:** In the post-survey, I measured perceived usefulness (PU), intention to use (ITU), and ease of use (PEOU) following the technology acceptance model of Venkatesh et al. (2003); Venkatesh and Bala (2008). Example items for the three constructs are: *"Imagine the tool was available in your next course, would you use it?", "The*

| Group | n | ITU | | PFA | | PASL | | PU | | PEOU | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| **TG: adaptive support** | 22 | 2.4259 | 0.5088 | 2.6666 | 0.5362 | 2.1388 | 0.7030 | 2.4722 | 0.5808 | 1.7407 | 0.6720 |
| **CG: non-adaptive support** | 23 | 2.9365 | 0.8667 | 3.3015 | 0.6984 | 2.6190 | 0.6690 | 2.9285 | 0.9122 | 2.0317 | 0.7951 |
| **Total** | 45 | *p = 0.0292** | | *p = 0.0027*** | | *p = 0.0364** | | *p = 0.0673* | | *p = 0.2243* | |
| | | t-value = 2.2797 | | t-value = -3.2066 | | t-value = -2.1741 | | t-value = -1.8889 | | t-value = -1.238 | |

Table 31: Measured mean and standard derivation on a 1-5 Likert scale (1: high, 5: low); *p <0.05, **p <0.01.
Source: Own illustration

*use of the tool enables me to write better argumentative texts."* or *"I would find the tool to be flexible to interact with"*. Moreover, I tested for the perceived feedback accuracy (PFA) (Podsakoff and Farh, 1989) of both versions of ArgueTutor to control for the manipulation of the adaptive feedback by asking three items: *"The feedback I received reflected my true performance.", "The tool accurately evaluated my performance."* and *"The feedback I received from the tool was an accurate evaluation of my performance"*. Also, I asked the participants to judge their perceived argumentation skill learning (PASL), following the model of Toulmin (Toulmin, 2003), by asking two items: *"I assume that the tool would help me improve my ability to argue in a structured way in texts."* and *"I assume that the tool would help me improve my ability to write more convincing texts."* The answers were captured on a 1-to 5-point Likert scale (1: totally agree to 5: totally disagree, with 3 being a neutral statement). Additionally, I asked three qualitative questions: *"What did you particularly like about the use of the argumentation tool?", "What else could be improved?"* and *"Do you have any other ideas?"*. Finally, I captured the demographics. In total, I asked 20 questions in the posttest.

#### 6.5.3.2 Results

For data analysis, I performed a double-sided t-test (Welch's t-test) to assess whether differences between both groups are statistically significant. In order to control for potential effects of interfering variables with my rather small sample size and to ensure that randomization was successful, I compared the differences in the means of the three constructs included in the pretest. For all three constructs, including personal innovativeness, feedback-seeking of individuals, and passive argumentative competency, I received p values larger than 0.05 between the treatment and the control group.

To investigate the impact of adaptive feedback in the conversational interaction of a conversational argumentation learning system, I compared the measured results between

the five constructs I captured in the posttest between the two groups. Participants of the TG rated the ITU, the PFA, and the PASL significantly higher than participants of the CG. A double-sided t-test confirmed that the participants of TG, who received adaptive argumentation tutoring by ArgueTutor, perceived the ITU, the PFA, and the PSAL of the pedagogical conversational agent statistically significantly better than participants who only received general tutoring support (see Table 31: ITU $p < 0.05$; PFA $p < 0.01$, PASL $p < 0.05$). This indicates that from a student's perspective the value of ArgueTutor not only lies in the dialogue-based interface and a step-by-step guidance but rather in the adaptivity of the pedagogical conversational agent. Moreover, the results indicate that the overall perception of ArgueTutor is very positive. In both groups, the ITU, the PASL, the PU, and the PEOU are higher than the neutral value of 3. Especially the perceived usefulness for writing better argumentative texts and the intention to use ArgueTutor as a writing support tool show promising results. A positive technology acceptance is especially important for learning tools to ensure students are perceiving the usage of the tool as helpful, useful, and easy to interact with. This will foster motivation and engagement to use the learning application. As described above, I also included open questions in this survey to receive the participants' opinions about the perception of the interaction of ArgueTutor to further evaluate my design functionalities and design principles. The general attitude of the interaction with ArgueTutor was very positive. Participants of the TG positively mentioned the fast and adaptive feedback (DP3 and DP4), the highlighting of argumentative components in their texts, and the adaptive feedback message on what to improve several times. In contrast, almost every participant of the CG mentioned that individual feedback on the written text with a personal message on what to improve is a necessary feature, indicating again the importance of adaptivity of a conversational argumentation learning system.

## 6.6 Discussion

In this study, I evaluate the impact of adaptive and intelligent feedback through the instantiated conversational argumentation learning system ArgueTutor. The experiment with 45 participants confirmed that the individual and adaptive feedback mechanism of ArgueTutor has a significant impact on participants' intention to use, perceived feedback accuracy of the pedagogical conversational agent, and subjective argumentation skill learning. The results indicate, that the novelty of conversational argumentation learning systems comes not only from the dialogue-based learning setting, but also the intelligent feedback in a user-centered and theory-motivated learning tool.

### 6.6.1 Theoretical and Practical Contributions

Therefore, the user-centered and theory-based design of an adaptive pedagogical conversational agent combined with intelligent algorithms interactively fosters learning according to the conceptual model of the ICAP framework (Chi and Wylie, 2014). Together with study 1, the evaluation of my instantiated conversational argumentation learning system ArgueTutor, confirms that adaptive dialogue-based tutoring helps students to increase their argumentation writing skills and also their motivation during learning compared to traditional argumentation scripting approaches (study 1) and non-adaptive conversational pedagogical agents (study 2).

Hence, the work makes several contributions to current research. To the best of my knowledge, this study is one of the first to present empirical insights into how to design a conversational argumentation learning system to foster argumentation skills of students based on adaptive and intelligent tutoring. It provides a foundation for researchers who also aim to develop learning tools to train metacognition skills to compare their solutions with mine. Educators can now use the design findings and principles to build their own adaptive pedagogical conversational agents to support argumentation learning in their large-scale or distance-learning scenarios.

### 6.6.2 Limitations and Future Work

My study also faces limitations. For the aim of this study, I focused the research on students from the university. Even though it is reasonable to assume that the transferability to other cases is possible without major changes, I cannot prove it with the research design. Moreover, even if 85% of the participants stated they have used a conversational agent before, novelty effects of students using the pedagogical conversational agent for the first time cannot be expelled in the empirical results. In these experiments, I prove the short-term influence of ArgueTutor on a student's argumentation skills and perceptions. For future work, I suggest measuring the long-term learning effects on students' skills. This can be achieved with a longitudinal study in a real-world learning setting, e.g., in tutoring the writing process of peer reviews in business model lectures.

Based on the qualitative user feedback, I see three main improvements to enhance ArgueTutor. First, I aim to improve the details of the argumentation text feedback by displaying the argumentative relations between the components through a graph engine. Therefore, students receive an additional overview of the discourse of their argumentation. Second, I aim to improve the transparency of the recommendations and explanations by providing a question mark button in the upper right corner of the conversational argu-

mentation learning system. Students can see transparent explanations for the adaptive tutoring at any time they want. Third, I want to train the conversational argumentation learning system on different argumentation mining corpora to extend the contribution of the learning tool to other domains and languages (e.g., English student essays or English law cases).

## 6.7 Conclusion

I this section, I designed, built, and evaluated ArgueTutor, an adaptive conversational argumentation learning system that individually tutors students with task explanations, theory input, guidance, and adaptive feedback on the argumentation structure of a text by leveraging the recent advances of argumentation mining algorithms.

I rigorously derived requirements from 85 scientific papers (rigor) and from twelve semi-structured user interviews (relevance) to formulate a concise set of design principles for adaptive pedagogical conversational agents for argumentation skill learning. I refined and evaluated those design principles as an instantiated artifact (ArgueTutor) in two empirical studies with 100 students in total.

Thereby, I answered RQ3b of my dissertation. The resulting design principles provide guidance for educators and designers when conceptualizing and implementing novel pedagogical conversational agents for metacognition skill learning. My results confirm that intelligent dialogue-based agents also impact argumentation skill learning of students.

The next section of my work will focus on transferring and evaluating the findings from section 5 and section 6 to another pedagogical scenario. This is necessary to not only reproduce the findings in other didactical domains but also to confirm or reject my findings from the previous sections.

# 7 Nudging Students Towards Self-Regulated Argumentation Learning

In this section of my dissertation, the objective is to (1) reproduce and transfer the findings on AI-based argumentation writing support systems to another pedagogical scenario and (2) to draw on the concept of digital nudging, to motivate learners to increase their argumentative writing based on argumentation self-evaluation. Thereby, I will answer RQ3c as stated in section 1.

**RQ3c:** *How can an argumentation writing support system be designed based on nudging and self-evaluation to influence students' argumentation writing skills?*

**Method:** *Design science research including corpus development, argumentation modeling in student-written business pitches, and mixed-method research with a series of experiments.*

**Results:** *Further empirical evidence for the influence of argumentation writing support systems based on certain design principles on students' argumentation skills and their learning experience.*

After investigating the influence of adaptive argumentation writing support on students' argumentation skills in section 5 (RQ3a) and the influence of dialogue-based tutoring on students' argumentation skills during a persuasive writing exercise in section 6 (RQ3b), this section aims to transfer the prior findings to another argumentative domain to ensure the validity of the results. Hence, I aim to build a novel learner-centered AI-based argumentation writing support system, which supports students with argumentation self-evaluation during writing a persuasive business pitch. To guide the design of a novel AI-based argumentation writing support system, I built on my findings from section 5 and section 6 as well as on the concept of digital nudging.

Embedding nudges in traditional learning environments, but also digital nudges in learning tools, has shown positive effects on learning outcomes, especially for self-regulated learning such as self-monitoring and self-evaluation (Caraban et al., 2019; Hartwig and Reuter, 2021; Weijers et al., 2020).

To investigate how individual argumentation self-evaluation will help students write more convincing texts, I design an argumentation writing support system called Argument-Feedback based on nudging theory and evaluated it in a series of three studies with a total of 108 students.

I found that students who receive a self-evaluation nudge wrote more convincing texts with a better quality of argumentation compared to two alternative groups. The measured self-efficacy and the technology acceptance provide promising results for embedding adaptive argumentation writing support tools in combination with digital nudging in traditional learning settings to foster self-regulated learning. [14]

[14]The insights presented in this section are partly based on my papers Wambsganss (2022) and Wambsganss and Niklaus (2022). I thank my collaborators for the valuable feedback on my work.

## 7.1 Introduction

The design and adoption of current argumentation learning systems is lagging behind the recent technological developments in human—computer interaction and NLP research. This eventually leads to a lack of literature and design knowledge from an interdisciplinary perspective for building an accessible, intelligent, and learner-centered argumentation tool that enables self-regulated learning (Xu et al., 2021). Thus, learners still struggle to train argumentation skills due to a lack of adaptive and instant evaluations in their individual learning process, since current approaches either fail to provide scalable and formative feedback on the argumentation skills of students or are not student centered (Lawrence and Reed, 2019; Scheuer, 2015). In the studies in section 5 and in section 6, I have successfully embedded argumentation mining algorithms in argumentation learning systems to support students when learning argumentation skills. However, even though much literature on argumentation systems exists, insights on the learner-centered system design in combination with argumentation mining as well as a transfer from peer reviews to other domains are rare.

To consider learner-centricity, I draw on the concept of nudging, originally a concept from behavioral economics (Thaler and Sunstein, 2009), to nudge users in their learning process. In fact, the soft-paternalistic paradigm of nudging users towards an intended behavior has shown positive effects on decision-making in digital environments (Caraban et al., 2019), e.g., for nudging towards secure internet behavior (Hartwig and Reuter, 2021; Zimmermann and Renaud, 2021). Nudging is based on psychological effects that influence or counteract behavior to support users in their decision-making, especially by exploiting biases and the heuristics of individuals (Mirsch et al., 2018). Embedding nudges in traditional learning environments and digital nudges in learning tools has shown positive effects on learning outcomes, especially for self-regulated learning such as self-monitoring and self-evaluation (Caraban et al., 2019; Hartwig and Reuter, 2021; Weijers et al., 2020; Motz et al., 2021). For example, Motz et al. (2021) have found that push notifications about missing submissions nudge students to improved assignment adherence and course grades, and Eskreis-Winkler et al. (2019) have found that nudging students to give motivational advice to their peers (e.g., how to stop procrastinating) leads to higher academic grades. However, even though literature on argumentation learning exists, extant research falls short of providing an ML-based approach combined with a student-centered and theory-based design along with rigorous empirical evaluations to compare distinct features such as a self-evaluation and nudging to help students learn how to argue in other scenarios than peer reviews. An interdisciplinary human—computer interaction perspective combined with recent methods from argumentation mining to form a learner-centered argumentation writing support system provides students with the opportunity

to monitor and evaluate themselves based on nudging theory is missing (Huang et al., 2016; Afrin et al., 2021; Wang et al., 2020).

Drawing on nudging theory as a guiding design paradigm (Thaler and Sunstein, 2009), I aimed to answer RQ3c by investigating, designing, and evaluating a novel student-centered learning system called ArgumentFeedback to enable students to self-monitor and self-evaluate their argumentation writing skills independently (Bjork et al., 2013; Bandura, 1991; Zimmerman and Schunk, 2001). I followed a rigorous theory-motivated design approach, where I systematically built on my findings from section 5 to carefully derive requirements for the design of ArgumentFeedback. Based on these design insights, I followed the methodology of Mirsch et al. (2018) to design different nudges for argumentation skill learning based on self-evaluation and self-monitoring (Cialdini and Goldstein, 2004). I iteratively evaluated the design with 108 students in three empirical studies to investigate if and how individual argumentation self-monitoring based on ML in combination with digital nudging helps students write more convincing texts.

To develop the AI-based argumentation writing support system, I (1) constructed a novel, theory-driven argumentation annotation scheme for student-written business model pitches, (2) compiled a corpus of 200 student-written texts, (3) trained a state-of-the-art model to predict the individual argumentation skill level, and (4) embedded the model in the student-centered and theory-informed design based on nudging theory (Weinmann et al., 2016; Mirsch et al., 2018). ArgumentFeedback provides students with individual argumentation self-evaluation by highlighting the argumentative components, by providing an adaptive learning dashboard for writing persuasive business pitches, and by adaptively nudging students to write more argumentative texts based on social comparison (Mirsch et al., 2017; Cialdini and Goldstein, 2004).

To determine the impact of adaptive self-evaluation with digital nudging for argumentative writing on students' argumentation skills, I evaluated the learning system in one qualitative study (study 1), one laboratory experiment (study 2), and one field experiment (study 3). I observed that participants who received an intelligent self-evaluation nudge including a social comparison and an adaptive feedback message when writing a persuasive business pitch wrote more formally argumentative texts compared to students who received an individual self-monitoring nudge and students who received only general self-monitoring. Furthermore, the perceived persuasiveness of the texts was significantly higher than of the texts from students using the other two tools. Moreover, the measured intention to use, the perceived ease of use, the perceived usefulness (Venkatesh and Bala, 2008), and the self-efficacy (Bandura, 1991) provide promising results for the usage of social nudging in combination with adaptive feedback for effective self-regulated argumentative skill learning. The results suggest that a student-centered learning tool

based on digital nudging combined with ML and NLP feedback helps students write more persuasive texts.

This work has three main contributions. First, I provide insights into a holistic design of a student-centered writing support tool based on recent advances in computational linguistics and digital nudging theory for self-regulated argumentation skill learning and metacognition skills in general. Second, I demonstrate the effectiveness of social nudging for argumentation learning by rigorously evaluating the tool ArgumentFeedback in three empirical studies with a total of 108 students. The results provide insights into the potential of designing student-centered learning tools based on NLP and ML in combination with digital nudging to foster self-regulated argumentative writing in a student's learning journey. Third, the research shows a case example of supporting metacognition skills in a scalable and individual format in possible large-scale scenarios. Thus, I provide design knowledge for other researchers and educators to design and compare similar tools for the support of metacognition skills of students based on computational methods and nudging theory. Hence, I foster the metacognition skill learning of students as a step to contribute to the OECD Learning Framework 2030 towards a metacognition-skill-based education.

## 7.2 Theoretical Background

I propose to combine recent advances in NLP, ML, and argumentation mining to evaluate new forms of human—computer interaction based on digital nudging theory for self-regulated argumentation skill learning (Bandura, 1991; Zimmerman and Schunk, 2001; Thaler and Sunstein, 2009). Following the suggested architecture of AI-based argumentation writing support systems in section 5, I not only address the educational task by providing feedback with recent computational methods but also focus on the learner through a user-centered and theory-motivated design to enable an adaptive and individual learning experience independent of an instructor, time, and place (Xu et al., 2021).

The complexity of using this ML, NLP, and argumentation mining in combination with digital nudging to provide students intelligent self-evaluation has, so far, been poorly assessed (Lawrence and Reed, 2019). In the approach, I focus on the first two subtasks to assess the argumentation level of a student to provide individual self-monitoring in combination with social nudging (Mirsch et al., 2018).

In 2009, Thaler and Sunstein (2009) introduced the concept of *nudging* to suggest guiding user behavior through the targeted design of user interfaces. They define a nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any option or significantly changing their economic incentives". The aim is

to build on the known systematic biases in human decision-making to support people in making better decisions (Caraban et al., 2020). Based on insights from behavioral economics, psychological effects that influence behavior are applied or counteracted to support users in their decision-making process (Thaler and Sunstein, 2009). The concept of influencing people with small design interventions to nudge them towards a certain behavior has been applied in several domains, such as information systems (e.g., Mirsch et al. (2018); Weinmann et al. (2016)) or human—computer interaction (e.g., Canca (2020); Lee et al. (2011); Acquisti et al. (2017)). More recently, the positive effects of nudging have been also used for the design of digital learning systems. Motz et al. (2021) have found that push notifications about missing submissions nudged students to improved assignment adherence and course grades compared to students who did not receive the nudge. Moreover, Eskreis-Winkler et al. (2019) have found that nudging students to give motivational advice to peers (e.g., how to stop procrastinating) leads to higher academic grades. However, to the best of my knowledge at the time of conducting this work, the concept of nudging has not been investigated for argumentation skill learning or argumentative writing support, e.g., to nudge users to write more persuasive texts.

I believe that embedding the concept of nudging for individual writing feedback could help students learn how to argue in a more effective way. I back up the hypothesis on self-regulated learning theory (Bjork et al., 2013). I believe that this theory supports the underlying hypothesis that individual and personal feedback including digital nudging on a student's argumentation motivates or engages the student to improve their skill level. Self-regulated learning theory reflects that students learn better with formative feedback and goal setting (Bandura, 1991; Zimmerman and Schunk, 2001). For students in a learning process, especially, critical reflection through self-monitoring and self-evaluation is an important component for effective learning (Zimmerman and Schunk, 2001). It can be an initial trigger for a student's learning process and thus the creation of new knowledge structures. However, the right portion of self-monitoring and self-reflection in combination with a learning goal is important for students to learn effectively (Bandura, 1991; Zimmerman and Schunk, 2001). The paradigm of nudging could provide guidance in better designing a self-evaluation mechanism (Canca, 2020), e.g., in how to *frame*, *anchor*, or *socially compare* the user's behavior to nudge them towards more persuasive writing (Thaler and Sunstein, 2009; Mirsch et al., 2018). For example, the concept of social comparison (Cialdini and Goldstein, 2004), also-called social nudge, might lead to a combination of learning feedback and goal setting (e.g., by stating a feedback message such as *"other users have used more arguments in the same task"*). For instance, nudges that include social comparisons with other learners have lead to better exam performance and an increase in study time (O'Connell and Lang, 2018). Theoretically, the underlying mechanism is also rooted in social cognitive theory (Bandura, 1977) and is thus, a prime

candidate for positive behavior change in learning processes (Bandura, 1991). Therefore, I aim to investigate if and how individual argumentation self-monitoring based on ML combined with digital nudging will help students write more convincing texts. To do so, I designed a novel pedagogical scenario based on a new annotated argumentation corpora of student-written texts to provide students adaptive self-evaluation in a real-world writing exercise.

## 7.3 Design and Development of ArgumentFeedback

In order to contribute to RQ3c on how to design an intelligent argumentation writing support tool with self-evaluation based on ML to influence students' argumentation writing skills, I built ArgumentFeedback. ArgumentFeedback is an AI-based argumentation writing support system and, thus, mainly based on my findings, the architecture, and the design principles from section 5. In this subsection, I will explain how I designed and developed the two main components of ArgumentFeedback: (1) the theory-based self-evaluation interface based on digital nudging theory and (2) the adaptive feedback algorithm to access the individual argumentation level of students. An overview of the basic concept of ArgumentFeedback is depicted in Figure 56.

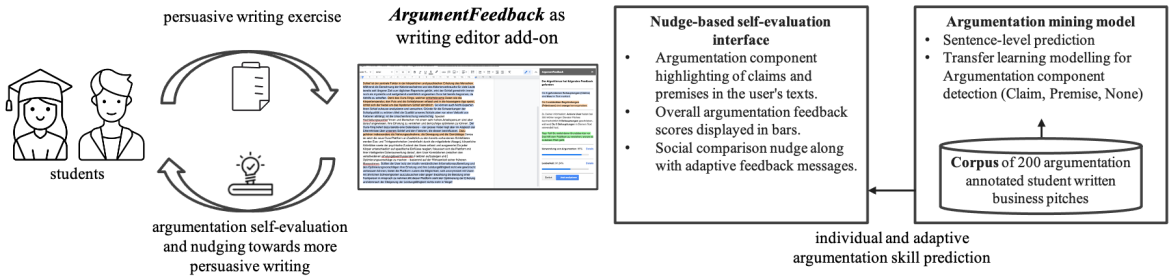



Figure 56: Basic concept of ArgumentFeedback: students are conducting a persuasive writing exercise in a common writing editor and are nudged with an intelligent self-evaluation towards more argumentative writing.
Source: Own illustration

### 7.3.1 Deriving Requirements for Nudging Students Towards more Persuasive Writing

Taking the results from my literature review for AI-based argumentation writing support systems and the conducted studies in section 5, my goal was to derive requirements and design knowledge for designing a novel nudge for argumentation skill learning. Based on self-regulated learning theory (Zimmerman and Schunk, 2001), the aim was to provide

students with self-monitoring and self-evaluation based on their current argumentation skill level. In order to achieve the goal, I followed the proposed methods for designing digital nudges by Schneider et al. (2018). The methods provide four practical steps on how to design a digital nudge for different scenarios, starting with (1) defining the goal, (2) understanding the users, (3) designing the nudge, and (4) testing the nudge.

#### 7.3.1.1 Defining Nudging Goal

The first step was to define the goal that was aimed to be achieved with digital nudging. My objective was to provide students with a tool that helped them improve their argumentation skills based on self-monitoring and self-evaluation independent of an instruction, time, and location. To quantify the desired user behavior, I aimed to design a feature that enabled users to write texts with a higher level of argumentation.

#### 7.3.1.2 Understanding the Users

Next, the decision behavior of the users was examined, and the psychological effects were investigated in the context of argumentation skill and digital nudge. Building on the derived theory-based design requirements on how to build an adaptive argumentation writing support system (see Table 32), the aim was to better understand the human heuristics and biases in the particular scenario. Hence, I performed a general literature review on digital nudging and digital nudges in educational settings specifically. I investigated the found literature under the criteria of (a) the underlying psychological effects (heuristics) of the intended user behavior, (b) the empirical proof of the effect, and (c) the insights on digital nudges to influence the user behavior. Following Mirsch et al. (2017); Caraban et al. (2019), I found seven heuristics to be suitable for the context of argumentation skill learning, namely *framing, status quo bias, social norms, loss aversion, anchoring and adjustment, hyperbolic discounting, decoupling*, and *priming*. After a more detailed evaluation and a workshop with two senior researchers who had much experience in user-interface design, digital nudging, and educational research, the concepts of *framing*, *social norms* and *anchoring and adjustment* were found to be most suitable for nudging students towards adaptive self-monitoring and self-evaluation based on recent methods from argumentation mining.

Along these lines, framing can be used to show specific consequences, implications, or eventualities of choices. For example, the technique of reinforcement can be used to strengthen the user's awareness of the desired behavior (Caraban et al., 2019). The

psychological effect of social norms can be applied to make comparisons with other people. In this regard, the nudging technique of social influence is used in literature (Caraban et al., 2019). Social nudging is based on the psychological effect of an individual's tendency to conform to the expectations of others (Cialdini and Goldstein, 2004). Hence, user behavior can be influenced by comparing one's performance to others. Finally, by nudging users through *anchoring and adjustment*, additional information is provided to students, which is perceived as a reference. Thus, decision-making can be facilitated, and users can be encouraged to perform a certain behavior (Caraban et al., 2019).

#### 7.3.1.3 Designing the Nudge

Based on these three derived concepts from literature, I built nine low-fidelity prototypes with different design instantiations of the three nudge groups (three design instantiations for each psychological effect). For example, for social nudging, I designed three different nudge instantiations for argumentation skill monitoring and evaluation. One nudge instantiation provided the users with a relative social comparison of their numbers of arguments (e.g., "80% of other users have used at least 6 arguments"). Moreover, users were provided with an absolute comparison by indicating how many more arguments others have written (e.g., "Other users have used two arguments more on average") and, thirdly, by a nudge indicating how many arguments others have written (e.g., "Others have used 8 arguments on average").

#### 7.3.1.4 Testing the Nudge

The different designs were then tested in qualitative evaluations with thirteen students from the university to receive feedback about the different designs (see study 1 in subsection 7.4). Based on these evaluations, I found that targeted information on the argumentative quality of their text facilitated the decisions of students to improve their written text argumentatively. Therefore, I implemented an informative framing nudge to provide students with self-monitoring. Second, the students stated that the social nudge was best suited to provide them with self-evaluation for argumentation skill learning. Most students mentioned that a social comparative feature based on the historical number of arguments other users had written for the same witting task in combination with an individual feedback message encouraged and motivated them the most to write more arguments. These two nudge designs were used in the evaluations later on (i.e., Figure 62).

| | **Design Requirement from Literature** | **Exemplary Guiding Literature** |
|---|---|---|
| 1) | Provide colored in-text argumentation highlighting to enable students to self-monitor their argumentation structures in their written text. | Soloway et al. (1994); Bjork et al. (2013); Zimmerman and Schunk (2001); Afrin et al. (2021); Chernodub et al. (2019) |
| 2) | Provide argumentation theory explanation and an argumentation learning goal (e.g., "write your texts as persuasive as possible") before and during the writing exercise. | Bandura (1991); Toulmin (2003); Soloway et al. (1994) |
| 3) | Provide summarizing argumentation scores in a learning dashboard with transparent explanations to enable students to receive a general overview of their argumentation skill level. | Soloway et al. (1994) |
| 4) | Provide students with adaptive feedback messages to guide them with feedback and self-evaluation based on their individual argumentation skill level. | Hattie and Timperley (2007); Wang et al. (2020); Vygotsky (1980) |

Table 32: Theory-based design requirements following my findings in section 5 on how to build an intelligent argumentation writing support system.
Source: Own illustration

### 7.3.2 User Interaction of ArgumentFeedback

Following the above-mentioned design requirements and the insights from nudging literature, ArgumentFeedback is built as a Google Doc's add-on that enables students to receive adaptive argumentation monitoring and evaluation whenever and wherever they want. A screenshot of ArgumentFeedback and the main functionalities (e.g., *F1 - F5*) can be regarded in Figure 57. The interface of ArgumentFeedback is kept rather simple to not distract users from the educational writing task. Hence, the learning tool is displayed on the right side of the text input field by providing adaptive argumentation self-monitoring and evaluation. When starting ArgumentFeedback, students first receive an introduction to the functionality of the tool. This ensures that first-time users are aware of the different functionalities of the writing support system. By referring to the Toulmin argumentation model, students are informed about the basics of argumentation theory (see F6 in Figure 57) (Toulmin, 2003). I integrated ArgumentFeedback in a popular writing editor. This comes with the advantage that users can not only write their texts in a familiar environment with several useful writing functions (such as spell check) but also then with ArgumentFeedback being thus more easy-to-access and easy-to-use on all kinds of advances through a responsive web-based design. In fact, to the best of my knowledge, with ArgumentFeedback, I present the first argumentation learning tool

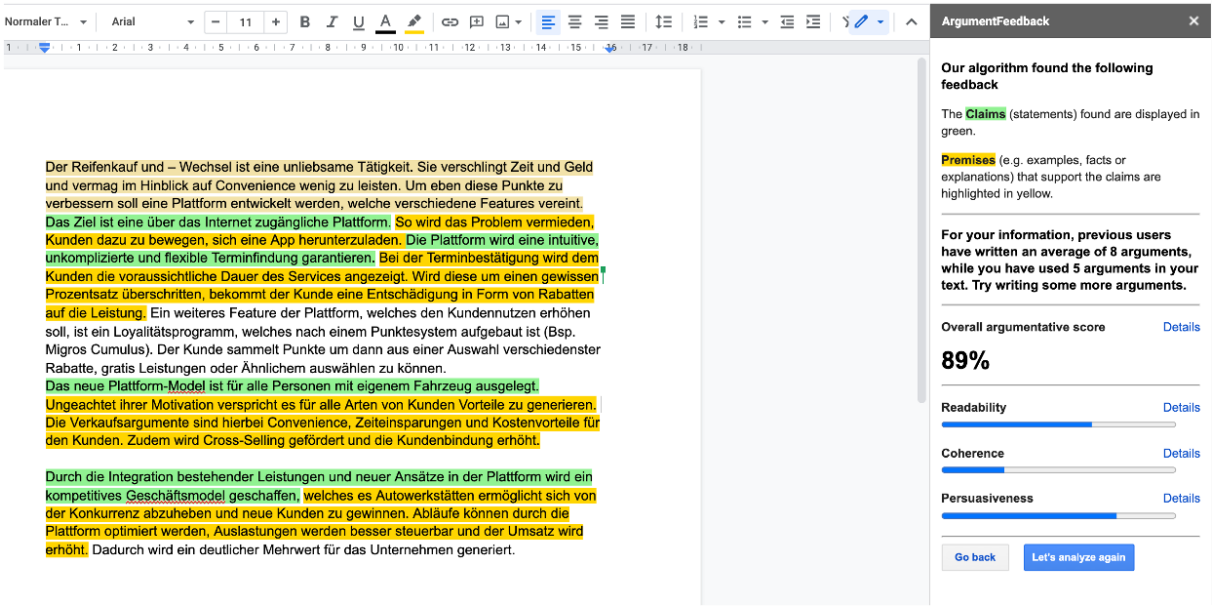



Figure 57: Screenshot of the AI-based argumentation writing support system ArgumentFeedback used in this study: a student conducts a persuasive writing exercise and is nudged with intelligent argumentation self-evaluation based on the argumentation quality of her text to increase augmentation skill learning.
Source: Own illustration

that is integrated into a common writing environment (such as Google Docs or Microsoft Office). Thus, I believe the availability as an add-on to a known cloud platform will foster its long-term usage and adaption compared to a stand-alone research tool.

After introducing the user to the function of the tool and providing a short overview of argumentation theory, ArgumentFeedback displays a simple self-evaluation dashboard to the user (Figure 57: F3, F4, and F5). After the user writes their text, they can select a certain text part and click on "now analyze". Afterwards, they will receive an adaptive argumentation evaluation based on their argumentation skills. To prevent information overload, I incorporated the function of only analyzing selected text paragraphs. The argumentation monitoring and evaluation consist of three different parts: (1) the argumentation component highlighting the claims and premises in the user's texts (F2), (2) the overall argumentation feedback scores displayed in bars (F5), and (3) the social comparison (nudge) along (F3) with the adaptive feedback message (F4). Through the argumentation highlighting, students can monitor their argumentation skill level. Claims are highlighted in blue and supporting premises in orange (see F2). Through color-coding, they can easily assess if a claim is supported or not and sufficiently backed up with a premise. This function provides the user with clear action steps on how to improve the persuasiveness and formal quality of their texts. Moreover, best practices and explanations about argumentation and argumentation theory are provided by clicking on the

"*Help*" button (see F8 in Figure 58).

The two summarizing scores (readability and persuasiveness) (*F5*) provide the student with a ranking of their text to provide general feedback. By clicking on the scores or "*details*", the methodology for calculating the scores, concrete hints, action steps, and explanations on how the student can increase their score level will be shown (see F7 in Figure 58). These action steps provide the user with orientation and context to improve their writing quality (Soloway et al., 1994; Hattie and Timperley, 2007).

The social nudge (F3 and F4) counts the number of claims in the user's document and provides a comparison based on historical data of the underlying knowledge base or task. In this case, I retrieved a corpus of 200 student-written business pitches. On average, students wrote eight arguments (claims) in a 300-word pitch. If students had more claims than the average, they received a positive feedback message (highlighted in green), such as: "Very nice. You displayed your idea with a sufficient amount of arguments." (see F4 in Figure 57). However, if a student had less than eight claims, he or she would receive a feedback message highlighted in yellow saying: "Maybe you could write some more arguments in your texts to improve the persuasiveness of your idea". The intelligence of ArgumentFeedback is based on a trained argumentation mining model.

To build an adaptive argumentation self-evaluation system based on recent computational methods, I collected and annotated a novel corpus to train and tune a model that provides students with adaptive argumentation self-monitoring and self-evaluation.

### 7.3.3 Annotating and Developing a Corpus of Argumentation-Annotated Student-Written Business Pitches

A major prerequisite for developing NLP methods that are able to identify argument components and argumentative relations in written texts is the availability of annotated corpora. As described in section 5, corpora that are applicable for the design and development of adaptive argumentative writing systems in pedagogical scenarios are rather scarce. In fact, at the time of conducting this study, only two corpus collections from the educational domain are available that are annotated for argumentative discourse structures. First, the English student essay corpus collection originated by Stab and Gurevych (2014a, 2017) and my German student-written peer review corpus explained in section 5. Consequently, there is still a *lack of corpora* for training models that provide students with adaptive feedback about the quality of their argumentation in different pedagogical scenarios, e.g., on persuasive business model pitches.

With this work, I propose a novel argumentation annotation scheme for persuasive student-

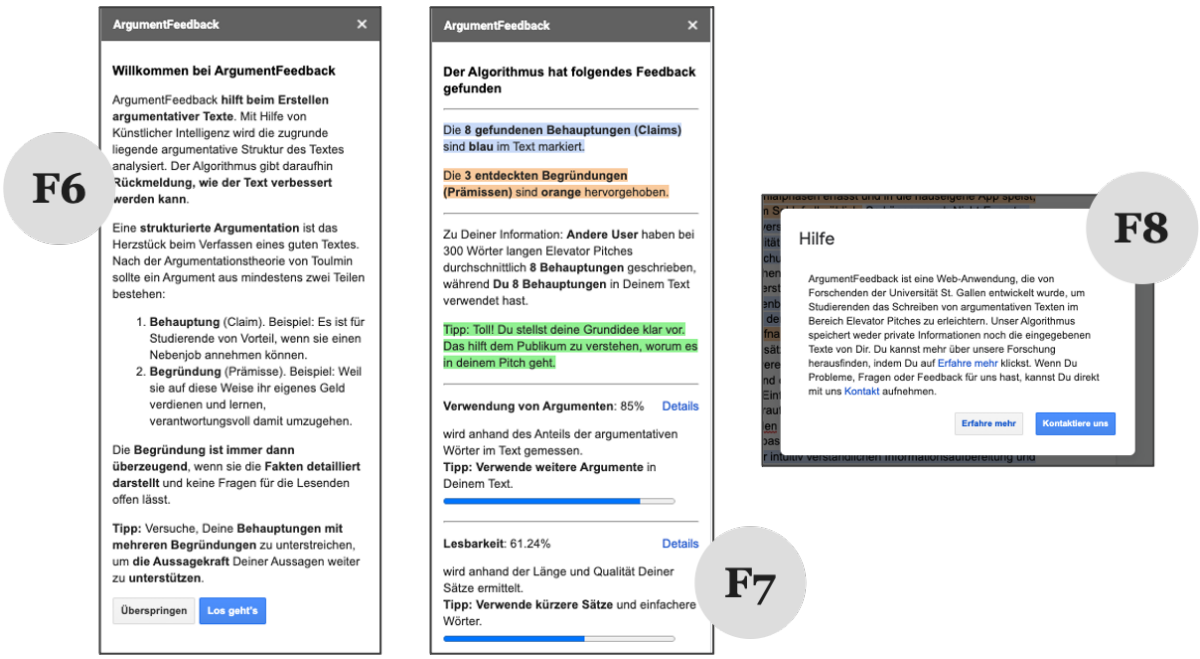



Figure 58: Additional functionalities of ArgumentFeedback. Left: argumentation theory introduction and goals setting based on the Toulmin model Middle: transparent and adaptive explanations of summarizing scores Right: help button for further information and learning support.
Source: Own illustration

written business model pitches. Therefore, I introduce a corpus of 200 student-written persuasive pitches with 3,207 sentences that are annotated for argument components, their relations, and persuasiveness scores to judge the argumentation quality of the single arguments.

#### 7.3.3.1 Introduction

My aim was to provide students with argumentation self-evaluation when writing a persuasive business pitch. However, no suitable corpus exists in the literature that is based on a rigorous annotation scheme, has a sufficient size to train a model with acceptable predictive accuracy, and fits the pedagogical scenario in German (Lawrence and Reed, 2019). Business model pitches – also called entrepreneurial or business pitches (Sabaj et al., 2020) – are described as *"a brief description of the value proposition of an idea or company"* (Daly and Davy, 2016) with the objective to convince a group of stakeholders of the novelty of an idea. A business model pitch usually lasts between 30 seconds and two minutes (Daly and Davy, 2016). The formulation of persuasive business model pitches is increasingly used in modern pedagogical scenarios, e.g., to train the entrepreneurship mindset or agile work (i.e., Oecd (2019)). Students are asked to write a concise but persuasive summary of the *"what, why, and how"* of their (business) idea in order to convince a peer. This pedagogical scenario is domain independent, easy to implement in different settings (e.g., in massive open online courses), and can be utilized to train skills such as logical argumentation. In fact, in their study about entrepreneurial business pitches, Fernández-Vázquez and Álvarez-Delgado (2019) found that "the lack of rational arguments determine the failure of the entrepreneur's efforts to be persuasive, regardless of the emotional appeals that are introduced into the pitch". Therefore, Fernández-Vázquez and Álvarez-Delgado (2019) calls for more emphasis on logical argumentation chains in business pitches.

However, linguistic research on business model pitches is a growing but still small field (Ducasse, 2020). Therefore, it is not surprising that no pitch corpus exists that is annotated for argumentation discourse structures based on appropriate argumentation schemes (Lawrence and Reed, 2019). I propose a new annotation scheme to model argument components in persuasive business model pitches. I based the annotation scheme on the model of Toulmin (1984) and the studies of Stab and Gurevych (2014a, 2017); Carlile et al. (2018); Ke et al. (2019).

#### 7.3.3.2 Corpus Construction

Following a 4-step methodology to build a robust corpus, I (1) searched the literature and scientific theory on argumentation discourse structures and argumentation models in different text domains; (2) randomly sampled 50 student-written business pitches and, based on my findings from step 1, developed a set of annotation guidelines consisting of rules and limitations on how to annotate argumentation discourse structures; (3) applied, evaluated and improved my guidelines with three native speakers in five consecutive workshops to resolve annotation ambiguities; and (4) applied the final annotation scheme based on my 26-page guideline to a corpus of 200 student-written business pitches with 3,207 annotated sentences.[15]

**Data Source**

I gathered a corpus of 200 student-written business model pitches in German. The data was collected in a mandatory business model innovation lecture at a Western European university. In this lecture, around 200 students develop and present a new business model. Students are asked to write a concise but persuasive pitch about the *"what, why, and how"* of their novel business idea in order to convince peer students. Afterwards, the students receive peer feedback from three fellow students on the persuasiveness of their business model pitch. The business pitches were collected from 2019 to 2020 according to the ethical guidelines of my university and with approval from the students to utilize the writings for scientific purposes.

**Annotation Scheme**

My objective is to model the argumentation discourse structures and the persuasiveness of student-written business model pitches by capturing argument components, their relations, and persuasiveness scores. The majority of the pitches in the corpus follow the same structure. They describe a novel business model and then provide convincing statements backed by examples, statistics, user-centered descriptions, quotes, or intuitions. However, I found that the specificity, strength, relevance, and evidence level vary between the different components. Thus, I captured them with qualitative labels on a 1-to-5 scale. The basic annotation scheme is illustrated in Figure 59.

**Argument Components**

For argument components, I follow established models in argumentation theory which provide detailed definitions of argument components (e.g., Toulmin (1984); Stab and Gurevych (2017)). These theories generally agree that a basic argument consists of multiple components and that it includes a *claim* that is supported or attacked by at least one *premise*. Also in student-written business model pitches, I found that a *claim* is the cen-

[15]The annotation guidelines can be found in the appendix of this dissertation.

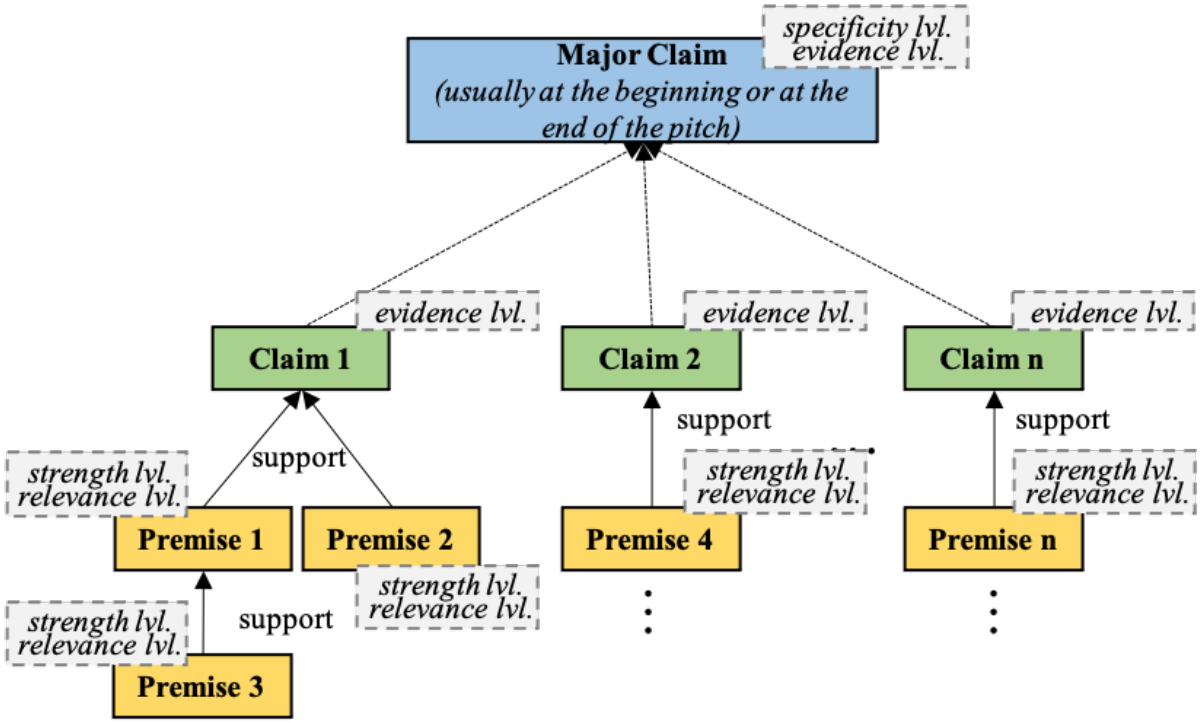



Figure 59: Overview of the argumentation annotation scheme for business model pitches, including argument components *(major claim, claim, premise)*, argumentative relations *(support)*, and persuasiveness scores *(specificity, evidence, strength, relevance)*.
Source: Own illustration

tral component of an argument. It is a controversial statement (e.g., claiming a strength or novelty of a business model) that is either true or false and should not be accepted by the stakeholder without additional support or backing. In business model pitches, authors usually start or conclude with an overall idea and topic of the business model. Similar to the persuasive student essays corpus by Stab and Gurevych (2017), I modelled this statement as a *major claim*. Usually, the major claim is present in the introduction or conclusion of the pitch - or in both. In the introduction, it often represents a general claim of the novelty of the business idea, whereas in the conclusion the major claim often summarizes or repeats the argumentation according to the author's business model idea. The major claim is then backed up by several other claims to manifest its validity. The *premise* supports the validity of the *claim* (e.g., by providing a statistic, analogy, user-centered example, or a value-based intuition). It is a reason given by the author to persuade the reader of their *claim*. Figure 60 illustrates a fully annotated example.[16]

### Persuasiveness Scores

To capture the differences in the persuasiveness levels of the components (i.e., the strength of a premise or the specificity of a major claim), I followed the approach of Carlile et al.

[16]Since the original texts are written in German, I translated the examples into English for the sake of this work.

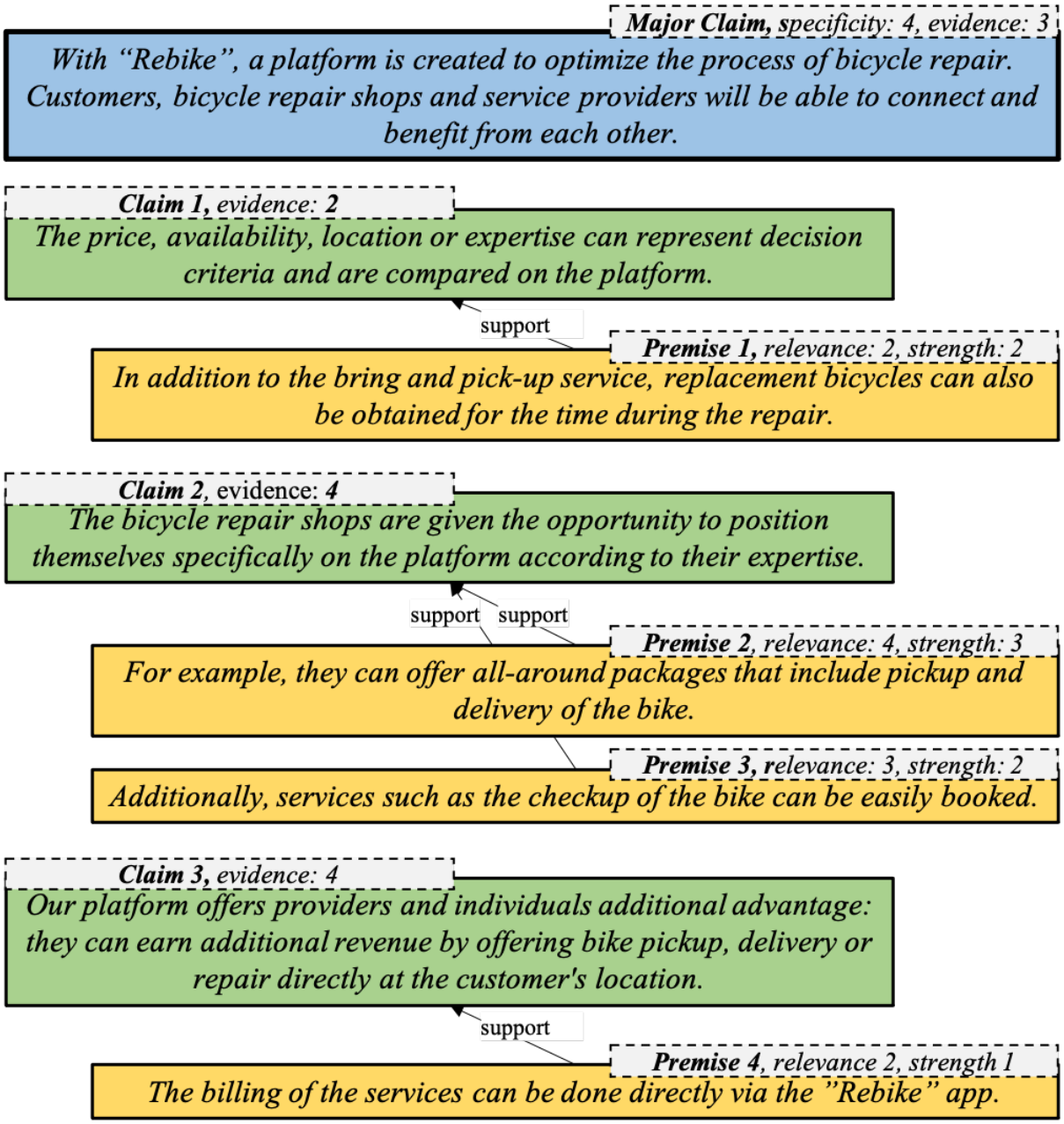



Figure 60: Annotated example of a pitch business model pitch according to the annotation scheme and guidelines.
Source: Own illustration

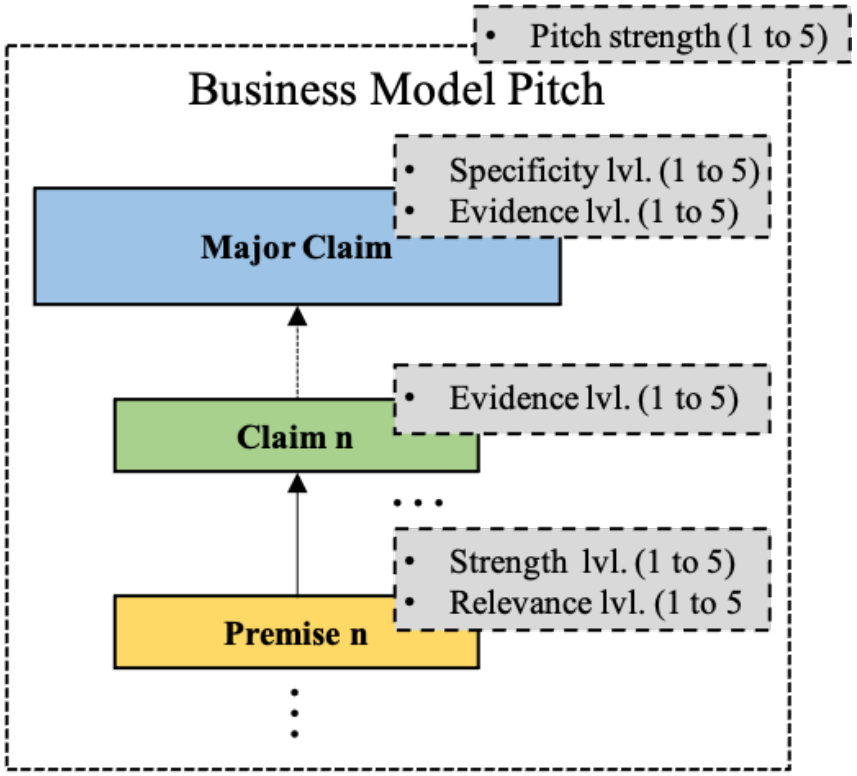



Figure 61: Overview of the annotation scheme including the argumentative components and the persuasiveness scores.
Source: Own illustration

(2018) and Ke et al. (2018) and defined five persuasiveness scores for the argumentative components. The objective was to capture the differences of a very persuasive major claim vs. a not very persuasive major claim accurately to provide students with more detailed writing support about *why* their argumentation is (un)persuasive. For the *major claim*, I found two attributes that differ in business model pitches: *specificity* and *evidence*. The *specificity* determines how detailed and specific the statement about the business model is, whereas the *evidence* ranks how well the major claim is backed up by supporting components. I found significant differences in both attributes throughout my corpus, which I aim to model with those scores. For *claims*, I defined *evidence* as a qualitative variable. Some claims seem to be strong in their statement. However, they do not contribute to the strength and persuasiveness of the overall business model. Thus, I specified *evidence* for a claim as the level of how well the claim supports the business model and/or the major claim. Most differences in the persuasiveness level in business model pitches can be found in the premises that back up the claims and thus the overall idea. I found premises to differ in two qualitative labels: *strength* and *relevance*. *Strength* is defined as how well a single premise contributes to the persuasiveness of the argument, and *relevance* determines how relevant a premise is for the overarching business idea. I believe that with these two scores I can model the most significant differences in the

| Score | Level/ components | Description |
|---|---|---|
| pitch strength | pitch | How argumentative and persuasive is the overall business model pitch? |
| specificity | major claim | How detailed and specific is the statement about the business model? |
| evidence | major claim | How well is the major claim backed up by supporting components? |
| evidence | claim | How well does the claim support the business model / the major claim? |
| strength | premise | How well does a single premise contribute to the persuasiveness of the argument? |
| relevance | premise | How relevant is a premise for the overarching business idea? |

Table 33: Summary of the descriptions of the persuasiveness scores on the different argumentative components.
Source: Own illustration

persuasiveness level of premises. Moreover, I found the business model pitches to also differ in their argumentative power on a discourse level. Sometimes a major claim is well formulated and supported by several claims and premises, but the business model is not really strong or novel in the overall picture because the argumentative discourse structure is weak. Therefore, I defined a document level score termed *"pitch strength"* to capture the persuasiveness of the argumentation discourse level of a business model. Table 34, Table 35, Table 36, Table 37 and Table 38 provide more detailed descriptions of the persuasiveness scores for the three argumentative components.

All qualitative attributes are measured on a 1-to-5 scale following Carlile et al. (2018), with every level being precisely defined in the annotation guidelines. A summary of the variables is illustrated in Table 33.[17]

**Annotation Process**

Two native German speakers annotated the business pitches independently from each other for the *major claim*, *claims* and *premises* as well as their *argumentative relationships*. Moreover, they labeled the *pitch strength*, the specificity and the evidence of the major claim, the evidence of claims, and the strength and relevance of premises according to the annotation guidelines I specified. Following my experience from the corpus annotation in section 5, my guideline consisted of 26 pages, including definitions and rules for what is an argument, which annotation scheme is to be used, and how argument compo-

[17]More elaborated definitions, examples, and keyword lists for both all attributes can be found in the annotation guidelines in the appendix of this dissertation.

| Score | Description |
|---|---|
| 5 | The major claim summarizes the argument well and has an addendum that indicates the extent to which the claim applies. Claims that summarize the argument must refer to most or all of the supporting components. |
| 4 | The major claim summarizes the argument very well by mentioning most or all of the supporting components. However, there is no addendum that states the conditions under which the claim is true. Alternatively, the claim moderately summarizes the argument by referring to a minority of the supporting components and includes an addendum. |
| 3 | The major claim contains a supporting component or addendum that indicates whether the claim is true. However, it does not adequately summarize the argument. |
| 2 | The major claim does not summarize the idea and does not contain an addendum that indicates whether the claim is true. |
| 1 | The major claim does not summarize the idea and is not explained by supporting components. It remains unclear to what the business model idea refers. |

Table 34: Description of the different *specificity* scores for major claims.
Source: Own illustration

| Score | Description |
|---|---|
| 5 | A very strong, very convincing argument. There are many supporting components that have high relevance scores. |
| 4 | A strong, persuasive reasoning pattern. There are enough supporting components with respectable relevance scores. |
| 3 | The reasoning pattern is present. However, the supporting components do not have high relevance scores. |
| 2 | A poor, only possibly persuasive reasoning pattern. There are a few supporting components. The relevance scores of the existing supporting components are low. |
| 1 | An unconvincing reasoning pattern. There are few or no supporting components. The relevance scores of the existing supporting components are low. |

Table 35: Description of the *evidence* score for major claims and claims.
Source: Own illustration

| Score | Description |
|---|---|
| 5 | The relationship between the premise and the claim is very clear. It is very easy to see how the premise contributes to the clarity and persuasiveness of the claim. |
| 4 | The relationship between premise and claim is clear. At least one of the components is very specific and clear, while the other component might be not specific. |
| 3 | The Relationship between premise and claim is only clear with imagination. It takes some thought to imagine how the components are related. Both statements refer to the same topic, but have no related ideas within the domain of the referred content. |
| 2 | The connection between premise and claim is not clearly evident. Some important assumptions are needed to relate the two components. A component may also receive this rating if both components have a low clarity. |
| 1 | The relationship between premise and claim is not apparent and is disjointed. Few people can see how the claim and premise are related. |

Table 36: Description of the *relevance* score for premises.
Source: Own illustration

| Score | Description |
|---|---|
| 5 | A strong premise. By itself, it contributes very well to the persuasiveness of the argument. |
| 4 | A reasonable premise. It is a fairly strong point, but it could be improved to increase its persuasiveness. |
| 3 | An inadequate premise. It is not a strong premise and may persuade only a few readers. |
| 2 | A weak premise. It can only help persuade a small number of readers. |
| 1 | The premise does not contribute to persuasiveness at all. |

Table 37: Description of the *strength* score for premises.
Source: Own illustration

| Score | Description |
|---|---|
| 5 | Little improvement or no improvement needed. The pitch describes, without any doubt, a very persuasive and strong business model. |
| 4 | The business idea is generally well understood, but can be expanded. |
| 3 | Poorly understandable idea due to errors or ambiguity. |
| 2 | It is unclear what idea the author wants to support argumentatively (no relevant idea, idea is incomprehensible). |
| 1 | The pitch does not introduce an idea. It remains totally unclear what the business model is about. |

Table 38: Description of the *strength* score for the pitch.
Source: Own illustration

nents, argumentative relations, and the qualitative attributes are to be judged. The entire guideline can be found in the appendix. After constructing the annotation guidelines, the results were discussed and validated by two independent senior researchers concerning the criteria of robustness, conciseness, extensibility, and comprehensibility. Several private training sessions and three team workshops were performed to resolve disagreements among the annotators and to reach a common understanding of the annotation guidelines. As done in section 5, I used the *tagtog* annotation tool[18], which provides a graphical interface for marking up text units and offers an environment for cloud-based annotation in a team. First, a text was classified into argumentative components (*major claim, claim, premise*) by the trained annotators. Second, the same annotators scored the argumentative relations and the qualitative attributes of the *major claim*, *premises*, and *claims* based on my annotation guideline on a 1-to-5 scale. After the first 50 pitches had been annotated by both annotators, we calculated the inter-annotator agreement scores. As we obtained satisfying results, we proceeded with a single annotator who marked up the remaining 150 documents.

#### 7.3.3.3 Corpus Analysis

I analyzed the results of the annotation process in order to examine the reliability of the corpus I built. In addition, I calculated some statistics of the final corpus. To evaluate the reliability of the argument component and argumentative relation annotations, I followed the approach of Stab and Gurevych (2014a).

**Inter-Annotator Agreement Argument Components**

With regard to the argument components, two strategies were used. Since there were no predefined markables, the annotators not only had to identify the *type of argument component* but also its *boundaries*. In order to assess the latter, we used Krippendorff's $\alpha_U$ (Krippendorff, 2004), which allows for assessing the reliability of an annotated corpus considering the differences in the markable boundaries. To evaluate the annotators' agreement in terms of the selected category of an argument component for a given sentence, I calculated percentage agreement and two chance-corrected measures, multi $\pi$ (Fleiss, 1971) and Krippendorff's $\alpha$ (Krippendorf, 1980). I decided to operate at sentence level since only 20.56% of the sentences in the corpus contain annotations of different argument components. Thus, evaluating the reliability at sentence level served as a good approximation of the inter-annotator agreement. At the level of individual sentences, 33.45% contain a claim, 35.64% a premise ,and 55.52% a non-annotation. 20.56% of the sentences

[18]https://www.tagtog.net/

| | % | Multi-$\pi$ | Krip. $\alpha$ | Krip. $\alpha_U$ |
|---|---|---|---|---|
| **Major claim** | 0.9948 | 0.9673 | 0.9673 | 0.5186 |
| **Claim** | 0.8729 | 0.7087 | 0.7088 | 0.5002 |
| **Premise** | 0.8768 | 0.7454 | 0.7455 | 0.5356 |
| **None** | 0.9261 | 0.7506 | 0.7506 | 0.2629 |

Table 39: Inter-annotator agreement of argument component annotations.
Source: Own illustration

| | % | Multi-$\pi$ | Krippendorff's $\alpha$ |
|---|---|---|---|
| **Support** | 0.9152 | 0.6139 | 0.6138 |

Table 40: Inter-annotator agreement of argumentative relation annotations.
Source: Own illustration

contain several annotations, with a combination of a premise and a non-argumentative span (32.21%) and a combination of a claim and a non-argumentative span (35.78%) representing the majority of those cases. Only 4.08% contain both a premise and a claim. At the token level, the following class distribution is achieved: 43.54% claim, 45.42% premise and 11.04% are not annotated.

Table 39 displays the resulting inter-annotator agreement scores. I obtain an inter-annotator agreement of 87.29% for the claims and 87.68% for the premises. The corresponding multi-$\pi$ scores are 0.71 and 0.75. Regarding Krippendorff's $\alpha$, a score of 0.71 and 0.75 is obtained, indicating a substantial agreement for both categories. With a score of 0.50 and 0.54, the unitized $\alpha$ of both the claim and premise annotations is somewhat smaller compared to the sentence-level agreement. Thus, the boundaries of argument components are less precisely identified in comparison to the classification into argument types. Yet the scores still suggest that there is a moderate level of agreement between the annotators. With a score of $\alpha_U$=0.26, the boundaries of non-argumentative units are less reliably detected. In contrast, the agreement scores targeting the component types are slightly higher for the non-argumentative spans as compared to the claims and premises (92.61%, 0.75, 0.75), indicating a substantial agreement between the annotators. Finally, with an inter-annotator agreement of 99.48% and a score of 0.97 for both multi-$\pi$ and Krippendorff's $\alpha$, I obtain an almost perfect agreement for the major claim components. Hence, I conclude that the annotation of the argument components in student-written business model pitches is reliably possible.

**Argumentative Relations**

To evaluate the reliability of the argumentative relations, I used the data set of all pairs of argument components that were possible during the annotation task according to the annotation scheme, i.e., all pairs of a major claim and a claim, a claim and a premise,

| | Cohen's $\kappa$ |
|---|---|
| **Strength of the pitch** | 0.8780 |
| **Strength of the premise** | 0.4660 |
| **Evidence of the major claim** | 0.3333 |
| **Evidence of the claim** | 0.2989 |
| **Specificity of the major claim** | 0.4110 |
| **Relevance of the premise** | 0.2760 |

Table 41: Inter-annotator agreement of the qualitative argumentation labels.
Source: Own illustration

and two premises. In total, the markables include 3,032 pairs of which 16.79% are annotated as support relations, while 83.21% of the possible pairs were left unidentified by an annotator. I obtained an inter-annotator agreement of 91.52% for the support relations. The corresponding multi-$\pi$ and Krippendorff's $\alpha$ scores both amount to 0.61 (see Table 40). Therefore, I conclude that argumentative relations can also be reliably annotated in business model pitches.

**Persuasiveness Scores**

Finally, I determined the reliability of the qualitative argumentation labels based on Cohen's $\kappa$ (Cohen, 1988). The results are displayed in Table 41. Considering the strength of the pitch, I obtained an almost perfect agreement between the two annotators ($\kappa$=0.88). With respect to the strength of the premise, I found moderate agreement ($\kappa$=0.47). The same applies to the specificity of the major claim ($\kappa$=0.41), which allows the conclusion that the annotators' labels are reliable. Regarding the evidence for both the claim and the major claim, as well as the relevance of the premise, there is some room for improvement. However, with scores of $\kappa$=0.33, $\kappa$=0.30, and $\kappa$=0.28, the annotations still show a fair agreement between the human annotators. Thus, qualitative argumentation labels can also be reliably annotated in business model pitches.

#### 7.3.3.4 Corpus Statistics

Table 42, Table 43 and Table 44 present some statistics of the final corpus. It consists of 200 student-written business pitches in German that are composed of 3,207 sentences with 61,964 tokens in total. Hence, on average, each document has 16 sentences and 305 tokens. A total of 262 major claims, 1,270 claims, and 1,481 premises were annotated. 1,069 textual spans were identified as not being an argument component ("None"). On top of that, 2,018 support relationships were marked up by the annotators. Table 45 presents some detailed statistics on the qualitative argumentation labels.

| | total | mean | stdev. | min | max | median |
|---|---|---|---|---|---|---|
| **Sentences** | 3,207 | 15.80 | 4.51 | 4 | 31 | 16 |
| **Tokens** | 61,964 | 305.24 | 67.12 | 104 | 473 | 317 |

Table 42: Distribution of *sentences* and *tokens* in the created corpus. Mean, standard deviation, minimum, maximum, and median refer to the number of sentences and tokens, respectively, per document.
Source: Own illustration

| | total | mean | stdev. | min / max | median | % |
|---|---|---|---|---|---|---|
| **Major cl.** | 262 | 1.29 | 0.70 | 0 / 4 | 1 | 6.42 |
| **Claim** | 1,270 | 6.26 | 2.05 | 1 / 13 | 6 | 31.11 |
| **Premise** | 1,481 | 7.30 | 2.46 | 2 / 14 | 7 | 36.28 |
| **None** | 1,069 | 5.27 | 3.24 | 0 / 17 | 5 | 26.19 |
| **All** | 4,082 | 17.87 | 5.21 | 3 / 34 | 18 | 100 |

Table 43: Overview on descriptive statistics for the types of *argument components*.
Source: Own illustration

Table 46 presents the distribution of support relationships in the corpus. With 46.30%, the majority of claims is not supported by any premise. 39.07% of the claims are backed up by exactly one premise, while 10.93% are supported by two premises. Patterns with more than two supporting premises rarely occur in the dataset. The majority of premises is not backed up by another premise (75.07%). However, in 21.08% of the cases, there is one supporting premise. In that way, more complex reasoning chains are established.

#### 7.3.3.5 Conclusion

The final corpus consists of 200 student-written business model pitches annotated for argument components, their relations, and six persuasiveness scores on different levels. Thereby, I contribute by deriving a detailed annotation scheme for a new data domain for argumentation mining based on argumentation theory and previous work on annotation schemes. Moreover, I present an annotation study based on 50 persuasive pitches that were annotated by two native German speakers, demonstrating that the annotation of student-written business model pitches is reliably possible. The resulting corpus (a) transfers my findings from section 5 in argumentation annotation of student-written text to the domain of business pitches and (b) serves as the underlying dataset for providing students adaptive argumentation self-evaluation in the course if this section.

| | total | mean | stdev. | min / max | median | % |
|---|---|---|---|---|---|---|
| **Support** | 2,018 | 5.47 | 1.76 | 1 / 11 | 5 | 12.75 |
| **None** | 13,814 | 72.51 | 41.06 | 2 / 242 | 72 | 87.25 |
| **All** | 15,832 | 77.99 | 42.66 | 3 / 253 | 78 | 100 |

Table 44: Types of *argumentative relations*.
Source: Own illustration

| | mean | stdev. | min | max | median |
|---|---|---|---|---|---|
| **Strength of the pitch** | 3.56 | 0.68 | 1 | 3 | 2 |
| **Strength of the premise** | 3.22 | 0.53 | 0 | 5 | 3 |
| **Evidence of the major cl.** | 3.71 | 0.46 | 3 | 4 | 4 |
| **Evidence of the claim** | 3.50 | 0.59 | 0 | 5 | 3 |
| **Specificity of the major cl.** | 3.58 | 0.51 | 2 | 4 | 4 |
| **Relevance of the premise** | 3.46 | 0.60 | 0 | 4 | 3 |

Table 45: Statistics on the qualitative labels.
Source: Own illustration

### 7.3.4 Modeling Argumentation Structures in Persuasive Business Pitches

After constructing and analyzing the corpus, I leveraged the novel data to train an ML model. My objective was to embed a classification algorithm in the back end of the argumentative writing support system to provide students with individual argumentation self-evaluation in the writing process. The task is considered a sentence-based classification task, where each sentence can be either a *claim*, a *premise*, or *non-argumentative* (major-claim labels are incorporated as "claims"). Therefore, I trained and tuned a long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) to classify the argumentative components of a given text. I tokenized the texts and transformed them into word embeddings. The data set was split into training and test sets using an 80:20 split. The LSTM architecture consisted of eight layers and a dropout rate of 0. For the component classification, I received an accuracy of 54.12%, a precision of 55.90%, and a recall of 54.12% on the test data. I benchmarked the approach against a BERT model (Devlin et al., 2018). For the BERT model, I found a learning rate of 1e-5, a batch size of 16, and training of the model of over 25 epochs to provide the best results. However, I received a rather unsatisfying accuracy of 47.50%, a precision of 46.66%, and a recall of 47.50%.

Furthermore, I calculated two summary scores to provide students with an overview of the quality of their argumentation based on previously extracted argumentative structures, including

- **Readability**: How readable is the text based on the Flesch Reading Ease score

| | #0 support | #1 support | #2 support | #3 support | #4 >= support |
|---|---|---|---|---|---|
| **Claim** | 0.4630 | 0.3907 | 0.1093 | 0.0230 | 0.014 |
| **Premise** | 0.7507 | 0.2108 | 0.0289 | 0.0073 | 0.0024 |

Table 46: Distribution of patterns of *support relationships* in the created corpus. The columns denote the percentage of claims and premises, respectively, that are supported by the specified number of premises.
Source: Own illustration

(Flesch, 1943)?

- **Argumentativeness**: How is the proportion of argumentative parts in the text compared to non-argumentative parts?

## 7.4 Demonstration and Evaluation

In this subsection, I describe the experimental evaluation and demonstration of the study. My goal was to answer RQ3c on how an individual argumentation self-evaluation tool based on ML can be designed to influence students' argumentation writing skills. Hence, I conducted three evaluation studies according to Venable et al. (2016). I (1) qualitatively evaluated the nudge design of the tool (study 1), (2) performed a proof-of-concept evaluation concerning the functionality and design of ArgumentFeedback, and (3) performed a proof of value field-experimental evaluation to investigate if and how individual argumentation self-evaluation in combination with social nudging and adaptive feedback will help students write more convincing texts. An overview of the experimental evaluation and the three studies can be found in Table 47. In total, I evaluated the tool with 108 students. All participants were bachelor's or master's students of business innovation and were recruited at a Western European German-speaking University. I ensured that no student participated in more than one study to control for intervention effects.

### 7.4.1 Study 1: Qualitative Design Evaluation

To ensure the validity of the design and incorporate the change requests from users about the functionality and the instantiations of the different nudges and pre-test effects for argumentation skill learning, I performed two studies (study 1 and study 2). The objectives of these studies were to verify if the theory-derived design requirements were of value to the students, to measure their completeness, and to identify student-based design requirements and requests.

#### 7.4.1.1 Study Design

The aim of study 1 was to evaluate the alpha version of ArgumentFeedback as a first prototype in Google Docs. My objective was to ensure that the design of the two nudges (argumentation self-monitoring and argumentation self-evaluation) lead to the intended effects and positive perceptions of the users. Therefore, I performed a qualitative evaluation, where I asked students to use the application to write a short persuasive paragraph of 50 to 100 words about a business idea. Afterwards, the students were interviewed about their experience with the tool and their perception of particular design features, such as the argumentation highlighting, the theory explanation, and the summarizing scores (to control the theory-based design requirements in Table 32). Moreover, participants were explicitly asked about the perception and the use of different digital nudges, which

| Study | Objective | Design | n | Results |
|---|---|---|---|---|
| Study 1 | - Ensure all theory-derived requirements are correctly addressed and instantiated in the design<br>- Ensure that the design of the two nudges (argumentation self-monitoring and argumentation self-evaluation) leads to the intended effects and positive perceptions of the users | Qualitative evaluation based on user observation and post-survey | 13 | - Fourteen change requests regarding the design and functionalities of ArgumentFeedback (e.g., concerning a "help" function, the size and color of the analyze button, and requests about the explanation and feedback texts of the two summarizing scores).<br>- Social comparative features based on the historical number of arguments in combination with an individual feedback message lead to high user satisfaction. |
| Study 2 | - Proof of concept evaluation of final version of ArgumentFeedback in a real-world educational writing scenario in an artificial and controlled set-up<br>- Ensure that technical artifacts provide high-quality argumentation self-evaluation | Laboratory experiment | 24 | - Perceived usefulness and perceived ease of use show promising results for using an intelligent writing support system based on the design for supporting students in writing argumentative texts (Venkatesh and Bala, 2008).<br>- The positive measured self-efficacy of the students might indicate that ArgumentFeedback increases self-regulated learning (i.e., Bandura (1991)). |
| Study 3 | - Empirically investigate the effects of intelligent argumentation self-evaluation in the field<br>- Evaluate to what extent an intelligent writing support system in combination with self-evaluation nudging influences the argumentation writing skills of students in a large-scale learning setting | Field experiment | 71 | - Intelligent argumentation self-evaluation including social comparison for students' argumentation skills on business pitches feedback helps them to write more convincing texts.<br>- Students using ArgumentFeedback including an intelligent self-evaluation nudge wrote texts with a better formal and perceived quality of argumentation to the ones receiving only a (intelligent) self-monitoring nudge. |

Table 47: Overview of the demonstration and experimental evaluation of ArgumentFeedback in three subsequent studies.
Source: Own illustration

were prototypically instantiated in different versions of ArgumentFeedback. For example, the participants were asked to indicate how they felt when they read the written social comparison information.

In total, thirteen students participated in study 1 (eight were graduate students, seven were undergraduates). Seven were male, and six were female with an average age of 24.0 (SD= 2.16).

#### 7.4.1.2 Results

Based on the qualitative answers, the observed usage, and the reported experience, several change requests for design instantiations were collected. For example, based on several user comments, the theory-explanations in ArgumentFeedback were shortened, and an introduction to the usage of the tool was added to provide the users with a quick overview of the core functionalities. Concerning the nudges, most students mentioned that the social comparative feature based on the historical number of arguments that the other users had written for the same witting task combined with an individual feedback message would encourage and motivate them the most to write more arguments. Hence, I incorporated that into the design and scripted different motivational messages based on the argumentation evaluation. Moreover, several smaller design requests were captured, such as including a "help" function, the size and color of the *analyze* button, and requests about the explanation and feedback texts of the two summarizing scores. In total, I captured fourteen change requests based on study 1 (five general requests, two about the introduction, two about the text explanations, three about the learning dashboard, and two about the digital nudges).

### 7.4.2 Study 2: Proof of Concept Laboratory Experiment

The aim of study 2 was to empirically evaluate the application of ArgumentFeedback in a real-world educational writing scenario. Hence, after I incorporated all of the change requests from study 1, I designed a laboratory experiment where students were asked to conduct a persuasive writing exercise and provided adaptive argumentation self-evaluation with social comparison and an adaptive feedback message (similar to TG2 in study 3 - see Figure 62).

In total, 24 students participated in this study. The participants were, on average, 24.31 years old (SD= 1.99); 11 were male, and 10 were female. 20 were studying at the master's level, and four were at the bachelor's level. All students were business students. I

controlled the experiment so that no student from study 1 also took part in study 2.

#### 7.4.2.1 Study Design

The experiment consisted of two parts: (1) an argumentative writing task and (2) a post-survey.

**(1) Argumentative writing task:** The experiment started with the students receiving a link to Google Docs, where they were asked to open the ArgumentFeedback add-on and become familiar with the tool. Next, they retrieved a business pitch of around 300 words. The students were asked to rewrite the pitch with the objective to increase its persuasiveness in order to convince a potential audience of investors. I asked all participants to use ArgumentFeedback to monitor and evaluate the argumentation level and the persuasiveness of their text. Providing the students with a pre-written text with incorporated argumentative flaws ensured that the students could focus on the argumentative part of the task and not on brainstorming their own novel business model. During the task, students received intelligent argumentation evaluation on the text, including a social comparison and an adaptive feedback message.

**(2) Post-purvey:** In the post-survey, I measured the perceived ease of use, the intention to use, and the perceived usefulness of the participants following the technology acceptance model of (Venkatesh et al., 2003; Venkatesh and Bala, 2008). Example items for the three constructs were: "*The use of the argumentation tool enables me to write better argumentative texts*" or "*I would find the tool to be flexible to interact with.*" Moreover, I measured the self-efficacy of students for the task of argumentation skill learning based on three items following (Bandura, 1991) to control for self-regulated learning. The items included, "*In comparison to other users, I will write a good argumentative text*", "*I am sure that I could write a very good argumentative text*", and "*I think I now know quite a bit about argumentative writing.*" Both constructs were measured with a 1- to 7-point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement). Also, I asked two questions to control for the digital nudge, such as "*I noticed on the analysis page the information on how many arguments I used compared to other users.*" and "*The information about how many arguments I used compared to other users motivated me to write more arguments.*". Finally, I captured the demographics and asked three qualitative questions: "*What did you particularly like about the use of the argumentation tool?*", "*What else could be improved?*" and "*Do you have any other ideas?*" In total, I asked 15 questions in the post-survey.

#### 7.4.2.2 Results

The *perceived usefulness* of students using ArgumentFeedback had a mean value of 4.65 (SD= 1.42). The *perceived ease of use* of the tool was rated with a mean value of 5.31 (SD= 1.81). All of the results were positive when compared to the midpoint scale of 4, indicating a positive technology acceptance of the writing support tool in the particular pedagogical scenario (Venkatesh and Bala, 2008; Venkatesh et al., 2003). Hence, I concluded that the design instantiations based on the derived theory requirements were satisfying for the user. Moreover, participants of study 2 judged the self-efficacy with a mean value of 4.62 (SD= 1.23). Following self-regulated learning theory, this might indicate that ArgumentFeedback could increase engagement and motivation through goal setting and adaptive feedback when practicing and learning argumentative writing (Bandura, 1991). Moreover, the two control questions for the evaluation of the social comparison nudge were judged with a mean of 5.94 (SD= 1.48), representing a positive result. Finally, I analyzed the qualitative answers and clustered similar responses into categories. In summary, the adaptive self-evaluation based on in-text highlighting and the summarized scores were positively mentioned several times. At the same time, students also mentioned that, sometimes, wrong argumentative components seemed to be highlighted by ArgumentFeedback. Some also asked for more details on the functionality of the tool, e.g., information about the accuracy of the feedback algorithm. I incorporated this qualitative feedback, e.g., by improving some transparent explanations about the functionality of the tool.

### 7.4.3 Study 3: Proof of Value Field Experiment

After evaluating the qualitative design of the approach (study 1) and conducting a proof of concept study with ArgumentFeedback (study 2), the objective was to empirically evaluate and investigate the effects of intelligent self-evaluation in the field. Hence, I designed a pedagogical scenario in which participants were asked to write a persuasive business pitch of 300 words. I manipulated the way students received adaptive self-monitoring and self-evaluation based on the persuasiveness of their pitches. Study 3 was conducted as a field experiment in a large-scale lecture about information management at a Western European University. Participants were randomly assigned to one of three groups, two treatment groups, and one control group. Students in the control group used a version of ArgumentFeedback which provided them with general argumentation self-monitoring based on syntactic rules (see Figure 56). Participants in treatment group one (TG1) used ArgumentFeedback with an adaptive self-monitoring nudge including argumentation in-text highlighting based on the trained ML model. Participants of the second treatment

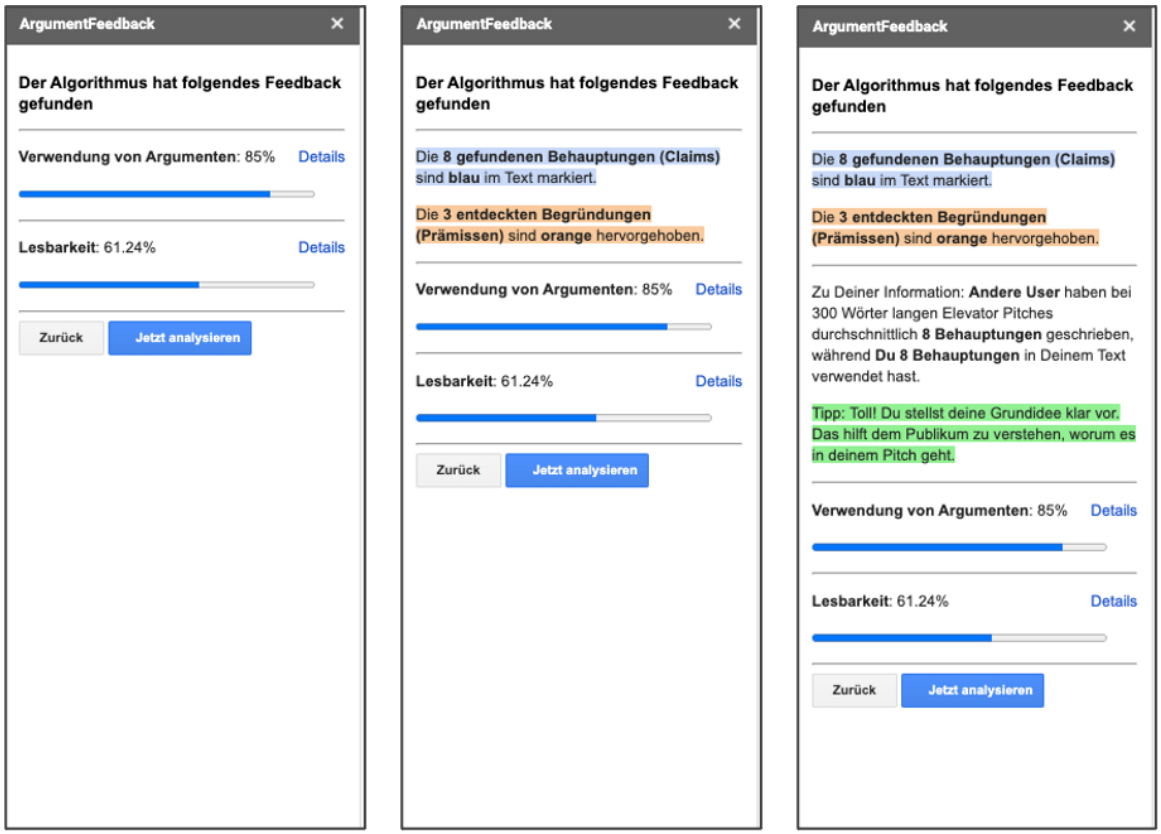



Figure 62: Overview of the three experimental groups: participants of the control group (CG) received argumentation self-monitoring based on syntactical rules while conducting an argumentative writing exercise. Students in treatment group 1 (TG1) conducted the same exercise and were nudged with intelligent argumentation self-monitoring based on the trained ML model. Participants in treatment group 2 (TG2) were nudged with an intelligent self-evaluation interface based on social comparison and an adaptive feedback message based on the trained ML model.
Source: Own illustration

group (TG2) wrote the persuasive business pitch by using ArgumentFeedback with an individual self-evaluation nudge based on social comparison and an adaptive feedback message based on the exact same trained ML model.

#### 7.4.3.1 Study Design

In the lecture I conducted the field experiment in, students developed and presented digital business models. In a particular assignment, the students were asked to submit a novel business idea based on a value proposition canvas (Osterwalder et al., 2014). Additionally, part of the assignment was to write a 300-word persuasive business pitch on the value proposition of their business idea. The stated task was to write a persuasive pitch in order to convince potential investors. Taking part in the assignment was not mandatory for students to pass the class, however, by successfully completing the assignment, the students could improve their final grade by 2.2%. The quality of the business pitch, e.g., the persuasiveness, had no influence on the evaluation of the assignment and, thus, no influence on final grades. I asked the students to write their business pitches on the Google Docs platform. Here, I provided them with different levels of feedback on the argumentation level of their business pitch. Nevertheless, I ensured that the students also had the opportunity to conduct the exercise in any other writing tool of their choice due to data privacy reasons and the ethical standards of the university. Students who chose to not use Google Docs in combination with ArgumentFeedback were excluded from the sample.

In total, 124 students from the course signed up for the assignment, from which 71 successfully completed the experiment. After randomization, I counted 21 valid results in the control group, 25 in TG1, and 25 in TG2. Participants of the control group had an average age of 24.17 (SD= 3.56, 9 males, 12 females); participants of TG1 were, on average, 24.24 years old (SD= 3.83, 13 males, 12 females); and participants in TG2 had an average age of 23.04 (SD= 2.32, 13 males, 12 females). The persuasive writing task took an average of 30 to 45 minutes. All participants of study 3 had neither participated in study 1 nor in study 2. The experiment consisted of three main parts: (1) pretest, (2) persuasive writing exercise, and (3) posttest. The pre-and post-phases were consistent for all participants. In the writing phase, TG1 used ArgumentFeedback with adaptive self-monitoring, and TG2 used ArgumentFeedback with adaptive self-evaluation based on social comparison and an individual feedback message to conduct the persuasive writing exercise, whereas participants of the control group conducted the same exercise using ArgumentFeedback with argumentation self-evaluation based on syntactic rules.

**(1) Pretest:** The experiment started with a pre-survey of eight questions. Similar to the evaluations in section 5 and in section 6, I tested two different constructs to assess whether the randomization was successful. First, I asked four questions to test the personal innovativeness in the domain of information technology of the participants following Agarwal and Karahanna (2000). The exemplary items were, *"I like to experiment with new information technologies,"* or *"If I heard about a new information technology, I would look for ways to experiment with it," "In general, I am hesitant to try out new information technologies,"* and *"Among my peers, I am usually the first to try out new information technologies."* Second, I tested the construct of feedback-seeking of the individuals following Ashford (1986). Example items were: "*It is important for me to receive feedback on my performance*" or "*I find feedback on my performance useful.*" Both constructs were measured with a 1- to 7-point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement).

**2) Persuasive writing exercise:** In the persuasive writing exercise of the experiments, I asked the participants to write a persuasive business pitch of 300 words. More information about the structure and the definition of a persuasive business pitch can be found in paragraph 7.3.3.1. Participants of TG1 used ArgumentFeedback with adaptive self-monitoring to write the pitch. Students in TG2 used ArgumentFeedback with adaptive self-evaluation based on a social comparison to conduct the same exercise, whereas participants of the control group conducted the same exercise using ArgumentFeedback with self-evaluation based on syntactic rules.

**3) Posttest:** Similar to study 2, in the post-survey, I measured the perceived ease of use, the intention to use, and the perceived usefulness of the participants following the technology acceptance model of Venkatesh et al. (2003); Venkatesh and Bala (2008) to ensure the validity of the design instantiation (exemplary item for intention to use: "*Imagine the tool would be available in your next course, would you use it?*"). Moreover, I again measured the self-efficacy of students for the task of argumentation skill learning based on three items following Bandura (1991) to control for self-regulated learning. All constructs were measured with a 1- to 7-point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement). Furthermore, I also asked students to self-monitor the argumentation quality of their texts. Hence, I included two items to capture (1) the self-reported number of arguments in the student's final business pitch and (2) how often they improved the argumentation level of their pitch based on the received feedback. Moreover, I incorporated one control question and two items for a manipulation check. Finally, I captured the demographics and asked, again, three qualitative questions: "*What did you particularly like about the use of the argumentation tool?*", "*What else could be improved?*", and "*Do you have any other ideas?*" In total, I asked 21 questions in the post-survey of study 3.

#### 7.4.3.2 Measurement of Argumentation Quality

Besides measuring several subjective constructs, the main objective was to measure the quality of the written texts from the three groups to investigate the research question. Therefore, I relied on my experience from the evaluation in section 5measuring two main variables: (1) the formal quality of the argumentation and (2) the perceived quality of the argumentation.

**(1) Formal quality of the argumentation**: The written business pitches were analyzed for their formal quality of argumentation. Similar as in section 5 and in section 6, I applied the annotation scheme for argumentative knowledge construction described by Weinberger and Fischer (2006). To measure the formal quality of the argumentation, the annotator had to distinguish between (a) *unsupported claims*, (b) *supported claims*, (c) *limited claims,* and (d) *supported and limited claims*. Therefore, I trained two annotators based on the 26-page annotation guidelines from the corpus to assess the argumentation components of the persuasive pitches. More information about the guidelines, the annotation study, and the inter-annotator agreements can be found in paragraph 7.3.3.2. The formal quality of the argumentation of the individual user was then defined by the number of arguments written by a user during the writing phase. Following Stegmann et al. (2012), only *supported*, *limited*, and *supported and limited claims* were counted as argumentation.

**(2) Perceived quality of the argumentation:** The perceived quality of the argumentation was annotated by the same two annotators. The objective was to subjectively judge how persuasive the given argumentation was on a Likert scale from 1 to 7 points (1: not very persuasive, 7: very persuasive). I took the mean of both annotators as a final variable for the formal and the perceived quality of the argumentation of the texts.

#### 7.4.3.3 Results

To answer the research question on how an intelligent argumentation writing support tool with self-evaluation based on ML to influence students' argumentation writing skills can be designed, the objective was to investigate the difference between the argumentation self-monitoring nudge (TG1), the argumentation self-evaluation nudge (TG2), and the control group on the students' argumentative writing skills and answer behavior. Moreover, I aimed to control for the design of the intelligent argumentation writing support tool and its impact on the students' perceived self-regulated learning.

Hence, I compared both the formal and perceived quality of the argumentation between the written texts of the control groups and the treatment groups. Moreover, I compared the self-reported argumentation quality and the argumentative changes of the students between all three groups. Therefore, I used an ANOVA analysis to check for significance between the variables and checked their assumptions visually with a test for normality and a test for homoscedasticity. All assumptions were met for all variables. To measure the effects between the groups, I ran a Tukey Post-Hoc test.

Finally, I compared the results of the perception measures, such as intention to use, perceived usefulness, and ease of use, and the self-efficacy to the midpoints scale to validate the design intervention of ArgumentFeedback for self-regulated learning.

To ensure that the randomization resulted in randomized groups and controlled for potential effects of interfering variables with the small sample size, I compared the differences in the means of the two constructs included in the pretest and the manipulation checks. For both constructs, including personal innovativeness and feedback-seeking of individuals, I received p values larger than 0.05 between the treatment and the control groups (for personal innovativeness p= 0.4697, for feedback-seeking of individuals p= 0.5047). This demonstrated that no significant difference in the mean values for these two constructs existed between the groups.

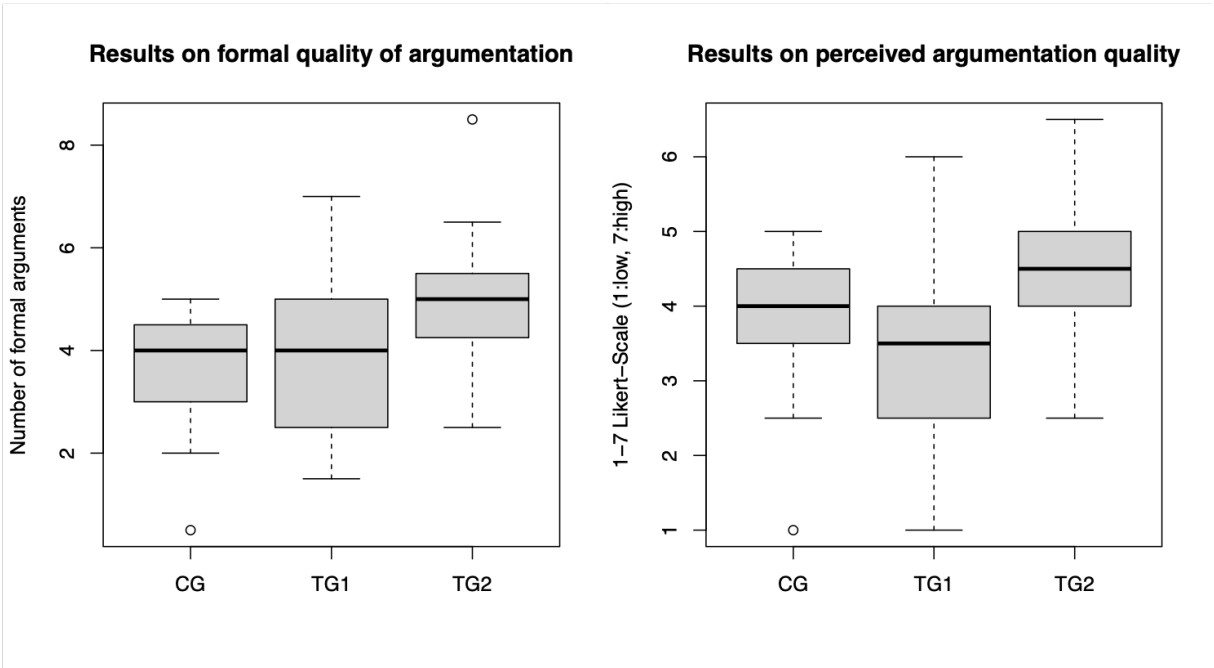



Figure 63: Results on formal (left) and perceived (right) quality of argumentation between the three groups (CG: argumentation self-monitoring, TG1: intelligent argumentation self-monitoring nudge, TG2: intelligent argumentation self-evaluation nudge).
Source: Own illustration

| Group | n | Mean | SD | p-value TukeyHSD Test TG1 | p-value TukeyHSD Test (CG) |
|---|---|---|---|---|---|
| **TG2: intelligent self-evaluation** | 25 | 4.875 | 1.2445 | 0.0478656* | 0.0130884* |
| **TG1: intelligent self-monitoring** | 25 | 3.9 | 1.6955 | - | 0.8124589 |
| **CG: self-monitoring** | 21 | 3.6428 | 1.2056 | - | - |

Table 48: Results of formal quality of argumentation between the groups on a 1-7 Likert scale (***p<0.001, **p<0.01, *p<0.05).
Source: Own illustration

| Group | n | Mean | SD | p-value TukeyHSD Test TG1 | p-value TukeyHSD Test (CG) |
|---|---|---|---|---|---|
| **TG2: intelligent self-evaluation** | 25 | 4.5 | 0.8597 | 0.0009243*** | 0.05159344 |
| **TG1: intelligent self-monitoring** | 25 | 3.34 | 1.3127 | - | 0.4242171 |
| **CG: self-monitoring** | 21 | 3.738 | 0.9568 | - | - |

Table 49: Results of perceived quality of argumentation between the groups on a 1-7 Likert scale, 1: not very persuasive, 7: very persuasive, 4: neutral (***p<0.001, **p<0.01, *p<0.05).
Source: Own illustration

**Formal argumentation writing quality**

The average number of arguments in the texts from the participants using ArgumentFeedback with intelligent self-evaluation (TG2) for the persuasive writing exercise was 4.875 (SD= 1.4). For the texts from the participants who received intelligent self-monitoring (TG1), I counted a mean of 3.9 arguments (SD= 1.69) (see Table 48 and Table 48). People in the control group wrote their texts with a mean of 3.64 arguments (SD= 1.20). The ANOVA analysis confirmed that the participants of TG2 wrote texts with a statistically significantly higher quality of formal argumentation: p= 0.0103 (p< 0.05). To measure the effects between the groups, I ran a Tukey Post-Hoc test to check for differences in the formal quality of the argumentation. The test confirmed that there was a significant difference between participants using ArgumentFeedback with an intelligent self-evaluation nudge (TG2) and participants receiving an intelligent self-monitoring nudge (TG1), with a p-value of 0.0478 (p< 0.05). Moreover, the test also confirmed students that TG2 wrote their text with a better formal quality of argumentation compared to the control group, with a p-value of 0.013 (p< 0.05). However, I did not find any significant difference between TG1 and CG (p= 0.8124). These results indicate that intelligent self-evaluation based on social comparison and an adaptive feedback message on the students' argumen-

tation in their business pitches helped them write more formally convincing texts.

**Perceived argumentation writing quality**

For the perceived quality of the argumentation (see Table 49), I found that, on a Likert scale from 1 to 7 points, (1: not very persuasive, 7: very persuasive, 4: neutral) texts from the TG2 achieved an average value of 4.5 (SD= 0.85). Participants of TG1 wrote their texts with a mean value of perceived quality of argumentation of 3.34 (SD= 1.31). For the CG, I measured a perceived quality of 3.73 (SD= 0.95). The ANOVA analysis and the Tukey HSD test showed that there was only a statistical difference between TG2 (intelligent self-evaluation nudge) and TG1 (intelligent self-monitoring nudge). No effects were found for the difference between the treatment and the control group. These results imply that self-evaluation based on social comparison and an adaptive feedback message leads to a higher perceived persuasiveness compared to only intelligent argumentation monitoring. However, the adaptivity of TG1 and TG2 did not yield a significant effect between the groups ($p<0.05$).

**Self-reported argumentation**

Next, I compared the self-reported argumentation and the reported argumentation changes between the three groups (see Figure 64 and Table 50). I found that students in TG2 reported 14.16 final arguments in their persuasive business pitch (SD= 5.63). This is significantly higher than what the students in the CG reported (mean of 8.0, SD= 4.11), with a p-value of 0.0036 ($p< 0.01$). Participants of TG1 reported a mean of 12.2 arguments (SD= 7.81). However, I did not find any significant difference between TG1 and TG2 or between TG1 and CG. Also, I compared the self-monitored argumentative changes that the participants reported between the groups (see Table 51). However, I could not find any significance between the distribution of the data. The results indicate that students who receive an intelligent self-evaluation nudge are self-aware about the higher number of arguments in their texts; however, they do not seem to change their argumentation based on the feedback more often.

**User perceptions and qualitative user feedback**

Finally, I evaluated the students' perception in order to check for the validity of the design instantiation of ArgumentFeedback. Thus, I calculated the mean across all of the groups for the perceived ease of use, perceived usefulness, and intention to use and compared them to the midpoints. All results were greater than the neutral value of 4, indicating a positive value for the design and the pedagogical scenario. A high perceived use fullness (mean= 4.2647, SD= 1.3227), intention to use (mean= 4.4975, SD= 1.2966), and ease of use (mean= 5.0096, SD= 1.1546) is especially important for learning tools to ensure

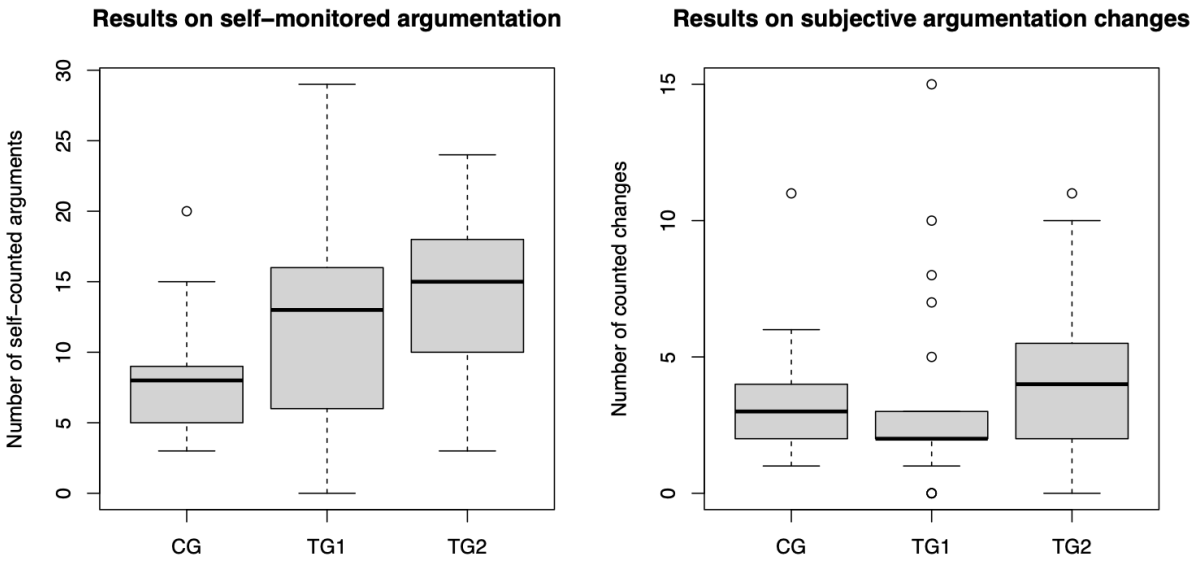



Figure 64: Results on self-reported argumentation of participants (left) and subjective argumentation changes of participants (right) between the three groups (CG: argumentation self-monitoring, TG1: intelligent argumentation self-monitoring nudge, TG2: intelligent argumentation self-evaluation nudge).
Source: Own illustration

| Group | Mean | SD | p-value TukeyHSD Test TG1 | p-value TukeyHSD Test (CG) |
|---|---|---|---|---|
| **TG2: intelligent self-evaluation** | 14.16 | 5.6311 | 0.5057191 | 0.0036889** |
| **TG1: intelligent self-monitoring** | 12.2 | 7.8155 | - | 0.0615984 |
| **CG: self-monitoring** | 8.0 | 4.1109 | - | - |

Table 50: Results of self-reported argumentation between the groups (***p<0.001, **p<0.01 *p <0.05).
Source: Own illustration

| Group | Mean | SD | p-value TukeyHSD Test TG1 | p-value TukeyHSD Test (CG) |
|---|---|---|---|---|
| **TG2: intelligent self-evaluation** | 4.08 | 3.18 | 0.6305731 | 0.6919273 |
| **TG1: intelligent self-monitoring** | 3.28 | 3.44 | - | 0.9980910 |
| **CG: self-monitoring** | 3.33 | 2.33 | - | - |

Table 51: Results of self-reported argumentative changes between the groups (***p<0.001, **p<0.01 *p<0.05).
Source: Own illustration

students are experiencing the usage of the tool as a benefit and that they find it easy to interact with. This will foster the motivation, engagement, and adoption of the learning application.

Moreover, positive effects for self-regulated learning can be also seen by comparing the means of the measured self-efficacy against the midpoints (Bandura, 1991). The average self-efficacy was 4.9338 (SD= 0.8644) on a 1-7 Likert scale. Compared to the neutral value of 4, this is a positive indication that argumentation self-monitoring, and self-evaluation help students learn in a self-regulated way.

I also asked open questions in the survey to receive the participants' opinions about the tool they used. The general attitude for ArgumentFeedback was quite positive. Participants positively mentioned the intelligent self-evaluation, the social comparison, and the integration of Google Docs several times. However, participants also asked if ArgumentFeedback could also provide concrete argument suggestions to provide more feedback on how to improve the argumentativeness. I translated the responses from German and categorized the most representative responses in Table 52.

## 7.5 Discussion

My results provide insights on the design and the impact of AI-based argumentation writing support systems on students' argumentation skills. The self-evaluation nudge, along with social comparison and feedback message on students' argumentation skills, helped the students write more persuasive texts compared to only monitoring their argumentative writing. Not only the perceived argumentation quality but also the formal quality of the argumentation was significantly higher for students receiving the intelligent self-evaluation compared to the ones receiving (intelligent) self-monitoring. Especially, the students with intelligent self-evaluation seemed to write with higher argumentation qual-

| Cluster | Exemplary User Response |
|---|---|
| General user interaction | *"ArgumentFeedback is simple and very easy to use. The integration in Google Docs, especially, makes the argumentation feedback very handy."* |
| Social comparison | *"The argumentative comparison with peers serves as an incentive to write at least as many or more arguments."* |
| Summarizing scores | *"Readability as a criterion is often neglected for argumentative texts. The feedback on this was very helpful for me to improve my pitch."* |
| Argumentative feedback | *"I was encouraged to write more arguments without stretching the text too much, as the ratio of the length of the text to the arguments is taken into account. Likewise, one is motivated to shorten the sentences. It is also good that tips are given for the categories in order to improve the evaluation."* |
| Dynamics of the interaction | *"In general, the tool should be more dynamic if possible. Going through an analysis every time is a bit cumbersome (tagging, etc.). Automatic updates of the feedback and metrics would improve the user experience a lot."* |
| Feedback accuracy | *"I'm not sure how well the tool recognized my arguments."* |

Table 52: Representative examples of qualitative user responses.
Source: Own illustration

ity compared to those who were only self-monitored. Both the formal and the perceived argumentation quality did significantly differ between both of these groups. Hence, to leverage computational methods for adaptive learning, self-evaluation nudging, including a peer comparison and an adaptive feedback message (as in TG2), seems to help students more compared to only monitoring their skill level through informative framing (as in CG or TG1).

Moreover, the self-reported items might provide an explanation for this phenomenon. The results indicate that students who receive an intelligent self-evaluation nudge are more self-aware about the higher number of arguments in their texts; however, they reported that the same number of argumentative corrections compared to the other groups. This implies that the argumentative changes implemented in TG2 seem to be more effective compared to the CG (i.e., self-reported argumentation divided by subjective changes - TG2: 3.47, TG1: 3.71, and CG: 2.42)

### 7.5.1 Theoretical and Practical Contributions

My results could be explained by nudging theory (Thaler and Sunstein, 2009). The social comparison seems to lead to a combination of learning feedback and goal setting (Cialdini and Goldstein, 2004). This is also reflected in the numerous qualitative comments I received about the peer comparison by user. Students seem to be nudged in their ar-

gumentative writing behavior, as social comparisons with other learners lead to better learning outcomes. This is also reflected in literature about social conformity (Cialdini and Goldstein, 2004). My study is in line with the results on social nudging for learning, e.g., with (O'Connell and Lang, 2018), who found that social comparison leads to higher exam performance and an increase in study time (O'Connell and Lang, 2018). The correct level of self-evaluation on a student's skills, such as argumentation skills, seems to lead to high self-efficacy and thus higher learning outcomes. Therefore, the findings contribute to self-regulated learning theory (Zimmerman and Schunk, 2001). My results seem to confirm the underlying mechanism rooted in social cognitive theory (Bandura, 1977) that intelligent self-evaluation with social comparison and adaptive monitoring leads to positive behavior changes in learning processes (Bandura, 1991). For argumentation skill learning, this result is especially novel. Past research on adaptive argumentation learning tools based on computational methods, such as (Afrin et al., 2021; Wang et al., 2020), have mostly focused on adaptive argumentation monitoring, e.g, based on in-text highlighting (Chernodub et al., 2019; Lippi and Torroni, 2016b; Afrin et al., 2021) or an argumentation discourse monitoring. Social comparisons for argumentative writing were mostly neglected or even showed that users did not prefer to compare themselves with peers. Moreover, traditional argumentation learning theory, such as the representational guidance theory by Suthers (2003), solely focuses on argumentation monitoring, e.g., supporting argumentation learning by providing representations of argumentation structures with the objective to stimulate and improve individual reasoning, collaboration, and learning (Pinkwart et al., 2009; Suthers and Hundhausen, 2001). My research provides novel insights on how to support students' argumentative writing based on self-evaluation and digital nudging more efficiently.

In order to successfully embed a writing support tool for students, e.g., in distance- or large-scale learning environments, and overall positive acceptance towards technology is very important to ensure the students perceive the usage of the tool as helpful, useful, and easy to interact with. This will foster the motivation and engagement in using the learning application. The positive ease of use, perceived usefulness, and intention to use of ArgumentFeedback by each group provide very promising results for the use of ArgumentFeedback as a self-monitoring and self-evaluation tool in different pedagogical scenarios.

Hence, the work makes several contributions, especially to human—computer interaction research and educational technology. I provide insights on a holistic design of a student-centered writing support tool based on recent advantages in computational linguistics and digital nudging theory for self-regulated argumentation skill learning and metacognition skills. Moreover, I demonstrate the effectiveness of social nudging for argumentation learning through rigorously evaluating the tool ArgumentFeedback in three empirical

studies with a total of 108 students. The results provide insight into the potential of designing student-centered learning tools based on NLP and ML combined with digital nudging to foster self-regulated argumentative writing in a student's learning journey. My research shows a case example of supporting metacognition skills in a scalable and individual way in possible large-scale scenarios. Thus, I provide design knowledge for other researchers and educators to design and compare similar tools for the support of metacognition skills of students based on computational methods and nudging theory.

### 7.5.2 Limitations and Future Work

Nevertheless, the work faces several limitations. First of all, the accuracy of the argumentation mining algorithm leaves space for improvement. I basically have two options to improve the performance of the underlying model: (1) by enriching the corpus with more annotated business pitches or (2) by improving the performance of the model. My objective is to improve the model by annotating the written business pitches that I collected in the experiment and adding them to the corpus. Second, the study was limited in sample size. Nevertheless, I believe certain effects where I investigated a trend, e.g., in the comparison of intelligent self-monitoring and general self-monitoring of self-reported argumentation ($p= 0.06165$), might result in a 95% significance level with further data. Finally, I only provided insights on the short-term effects of the design interventions. In the experiment, I showed that self-evaluation has a short-term influence on a student's argumentation skills. For future work, I suggest measuring the long-term learning effects on students' skills based on different nudging interventions.

Finally, I want to highlight the ethical limitations of the study. Regarding the implementation of the intelligent writing support system, I do not want to replace human educators by any means, as I believe that skilled teachers will always be able to provide better individual writing support than ArgumentFeedback. However, I hope that through the system, human educators can focus more on detailed questions and devote more time to difficult cases. Also, I see several data privacy concerns about integrating the tool in a common writing editor, since personal data, e.g., about the argumentation skill level of students, might be exposed to large stock-based tech companies. Hence, I call for future discussions on how to overcome the trade-off of making novel user-centered learning applications widely accessible and easy to use, e.g., by integrating them in known cloud environments without exposing learner data to third parties.

## 7.6 Conclusion

In this section, I answered RQ3c by (1) constructing a novel, theory-driven argumentation annotation scheme for student-written business model pitches, (2) compiling a corpus of 200 student-written texts, (3) training a state-of-the-art model to predict the individual argumentation skill level, and (4) embedding the model in a student-centered and theory-informed design based on nudging theory. I evaluated the final tool, ArgumentFeedback, in three studies and observed that participants who received an intelligent self-evaluation nudge including social comparison and an adaptive feedback message when writing a persuasive business pitch wrote more formally argumentative texts compared to students who received an individual self-monitoring nudge and students who received only general self-monitoring. The measured intention to use, the perceived ease of use, the perceived usefulness (Venkatesh and Bala, 2008), and the self-efficacy (Bandura, 1991) provide promising results for the usage of social nudging in combination with adaptive feedback for effective self-regulated argumentative skill learning. The results suggest that a student-centered learning tool based on digital nudging combined with ML and NLP feedback helps students write more persuasive texts.

After investigating the influence of adaptive argumentation writing support on students' argumentation skills in section 5 (RQ3a) and the influence of dialogue-based tutoring on students' argumentation skills during a persuasive writing exercise in section 6 (RQ3b), this section contributes to this dissertation by transferring the prior findings to another argumentative domain to ensure the validity of the results.

I evaluate and reproduce the findings from section 5 and section 6 of my dissertation by investigating the impact of AI-based argumentation writing support systems in (a) another pedagogical scenario (persuasive business pitches) and (b) another design paradigm (digital nudging for self-regulated argumentation learning).

# 8 Summary of Contributions and Areas for Future Research

The support for individuals with technology-mediated tools to learn and practice argumentation is a growing and important research field for a metacognition-based education (WEF, 2018; Oecd, 2018). Hence, the main objective of this dissertation project was to contribute to three outlined research challenges in the fields of NLP, human—computer interaction, and IS research. To fulfill this overall objective, the thesis was organized concerning the investigation of two overarching research areas, which relate to (1) *how to design novel argumentation learning systems based on recent advantages of AI to support argumentation skill learning irrespective of instructor, time, and location*, and (2) *how these design findings influence students' argumentation skills and learner experience.*

In this section, I address the contributions and consequences of my research, particularly in relation to the literature on argumentation learning systems, argumentation mining, and metacognition skill learning in general. While I have already discussed the individual studies in other areas of my dissertation, the purpose of this section is to discuss how the various studies handled the research problems I outlined in section 1. As a result, this part presents a unified overview of the contributions of the separate research to theory and practice. I conclude my work by proposing potential future research areas that are not included in this dissertation.

## 8.1 Theoretical Contributions of the Dissertation

### 8.1.1 Design of NLP-based Adaptive Educational Learning Tools

The first research question of the dissertation focused on the technical foundations of NLP and ML for adaptive learning tools as well as on the requirements for argumentation learning systems:

**RQ1:** *What are the foundations and requirements for adaptive learning tools to teach students how to argue?*

In section 4, I analyzed the theoretically grounded and empirically validated dimensions and characteristics of NLP features for ML-based IS and proposed a novel design space of features and categories incorporated in an easily usable feature canvas to aid both researchers and practitioners in approaching NLP endeavors in a structured manner. The

NLP feature taxonomy aids educational designers in the development, refinement, comparison, and evaluation of their NLP pipelines, such as when creating AI-based educational applications. As a result, I not only demonstrated the diversity of text feature engineering in diverse application domains by a systematic literature analysis, but I also described a classification framework using a taxonomy with five dimensions and several attributes. The resulting taxonomy helps to compare NLP features across research, domains, and applications, as well as reduces the resources spent by practitioners and scientists on feature engineering using traditional machine learning approaches.

I integrate existing research from a theoretical standpoint by creating a new taxonomy that goes beyond existing categories and structures to group NLP features from an integrative IS perspective. I present a greater knowledge of what has to be considered while creating, assessing, and comparing NLP applications using a common taxonomy of NLP features taking a holistic perspective. The resulting taxonomy better assists educational designers in determining which design characteristics influence a specific outcome by merging the dimensions and characteristics of my taxonomy (e.g., for explainability and transparency post-hoc analysis of educational tools). This opens the door to further research and more in-depth understanding of how to build, evaluate, and compare NLP features. The taxonomy does more than classify NLP-based features; it also facilitates knowledge exchange, improves understanding of interrelationships among NLP features, and thus aids researchers and practitioners in their decision-making (Bailey, 1994; Kundisch et al., 2021; Usman et al., 2017). Furthermore, the numerous characteristics are classified and grouped into three clusters (C1, C2, C3). These clusters will aid scholars and practitioners in gaining an overview of present NLP feature engineering research and identifying gaps in the field. My research findings, I believe, can serve as a useful starting point for further developing NLP feature engineering guidelines or frameworks, particularly for special use cases where dealing with accuracy issues in small data sets or conducting explainability assessments of large data models are critical for successful use case deployment. This provides a strong foundation for technology-mediated learning, such as argumentation writing support systems or conversational argumentation learning systems, to better train NLP models for ML-based learning tools. The taxonomy can also be used to teach and train novice NLP learners by providing a comprehensive classification of NLP features across domains and grouped clusters of applications.

In subsection 4.2, I followed the insights from the developed NLP features taxonomy and took a deep dive into the field of argumentation mining. Based on a systematic literature review, I built a domain-agnostic model for argumentation identification, an approach that has been neglected so far at the time of conducting the study. In this vein, I contribute to scientific literature in argumentation mining by demonstrating a new modeling approach and solution to current problems of argumentation mining. The proposed

model demonstrates that transformer-based models might help to decrease the amount of needed feature engineering for argumentation mining algorithms and might contribute to domain-agnostic classification approaches for argumentative texts. Finally, applying the power of transfer learning models like BERT to argumentation mining research may give a solution to the fragmentation of research into various domain-specific nodes.

Both the corpus-diagnostic algorithm and the NLP feature taxonomy assist educational designs to build artifacts based on argumentation mining, e.g., argumentation learning systems. The NLP feature taxonomy and the algorithm provide a solid theoretical and practical foundation to further impact the design of adaptive learning tools for argumentation skills.

### 8.1.2 Design of Adaptive Argumentation Learning Systems

The main contribution of my work focuses on the design, development, and evaluation of argumentation learning systems. While research has recognized AI-based argumentation learning systems' potential for supporting students and educators and possibly transforming institutional activities, such as persuasive essay scoring (Afrin et al., 2021; Scheuer, 2015; Scheuer et al., 2010; Stab and Gurevych, 2017), little attention has been paid to the design, development, and evaluation of IS based on AI capabilities (i.e., argumentation mining) and socio-technical system embedding (i.e., in technology-mediated learning). This work combines previous findings and theories for the design of technology-mediated learning systems and the state of the art of argumentation mining technology (Afrin et al., 2021; Bostrom and Heinen, 1977; Fan et al., 2020; Huang et al., 2021; Scheuer et al., 2010) to provide comprehensive guidelines and principles for developing and implementing AI-based argumentation learning systems in higher education (Matook and Brown, 2017). The effective design and implementation of AI-based argumentation learning systems are valuable not only for students learning argumentation skills and educators judging students' argumentation levels but also for higher education institutions operating in an increasingly complex and competitive environment (Daniel, 2015). My dissertation introduces and contributes to two novel classes of argumentation learning systems: AI-based argumentation writing support systems and conversational argumentation learning systems.

Both systems classes address RQ1, RQ2, and RQ3 in my dissertation:

**RQ2:** *How can an adaptive learning tool be designed based on AI in a large-scale pedagogical scenario to support the development of students' argumentation skills in their learning*

*process?*

**RQ3:** *To what extent does an AI-based adaptive learning tool influence the argumentation skills of students in different pedagogical scenarios?*

Although the novel systems class may be considered part of other educational IS types, such as intelligent tutoring systems (Kulik and Fletcher, 2016) or adaptive argumentation support approaches (Scheuer, 2015), their architecture differs fundamentally from existing systems in their accessibility, intelligence, and socio-technical and user-centered design (see section 2).

Educational theories in general (such as self-regulated learning theory) and argumentation learning theories specifically (i.e., Bjork et al. (2013); Lawrence and Reed (2019); Metcalfe (2017); Scheuer et al. (2010)) demonstrate the potential to investigate AI-based argumentation writing support systems and conversational argumentation learning systems as a new genre of educational IS to provide self-evaluation for the maximum impact of IS in educational settings. Therefore, to answer RQ1 and RQ2, I derived and evaluated a set of design principles for developing and implementing AI-based argumentation writing support systems and conversational argumentation learning systems in different pedagogical scenarios.

Building on self-regulated learning theory as a guiding kernel theory (Bandura, 1991; Zimmerman and Schunk, 2001), I derived four design principles in section 5 for AI-based argumentation writing support systems and instantiated, demonstrated, and evaluated them by implementing this IS type in a higher education context (Matook and Brown, 2017). Moreover, I reproduced and validated my findings on the impact of AI-based argumentation writing support systems in the area of student peer reviews and in the domain of persuasive business model pitches (see section 7). The results from both sections indicate that highlighting learners' logical argumentation errors in combination with a self-evaluation dashboard helps them to foster their formal metacognitive argumentation skills. Furthermore, I investigated the short-term effects of AI-based argumentation writing support systems based on my design principles on students' perceived argumentation skills within the pedagogical domain of the task.

With respect to conversational argumentation learning systems, in section section 6, I investigated, instantiated, demonstrated, and evaluated the design of dialogue-based argumentation tutoring. The results of two empirical laboratory studies indicate that individual argumentation tutoring in combination with formative feedback helps students to

increase their formal and perceived argumentation skills.

Learners stand to benefit from both system classes and the application of AI-based argumentation learning systems as well as educators and educational institutions, which could increase their efficiency and effectiveness. For example, AI-based argumentation writing support systems could also be utilized to automate argumentative essay correction on a deeper granular level (sub-sentence-based). Educators may rely on an AI-based argumentation writing support system to provide reliable support in scoring student-written essays' argumentation at scale (Horbach et al., 2018). Specifically, the more theory-based view of AI-based argumentation writing support systems as a new system class for technology-mediated learning can encourage educational IS researchers to further investigate, implement, and drive the research and functionalities of AI-based argumentation learning systems. The systematic conceptualization of underlying design principles for AI-based argumentation writing support systems provides a solid foundation for further development and implementation of such systems. In particular, the service-oriented architecture perspective of AI-based argumentation writing support systems (see section 5) can encourage research in different domains and disciplines to contribute to a metacognition (argumentation) skill-based education in the future, (e.g., for computer linguistics by providing more student-written corpora and models or for human—computer interaction by providing investigated interaction designs for novel metacognition skill training).

Moreover, I exemplify how to embed recent argumentation mining technology in an IS context. Argumentation mining is a powerful method and toolset for investigating and contributing to various IS research phenomena (e.g., in the context of opinion mining, decision support systems, user-generated content, or social media analysis — for reviews on argumentation mining please refer to Lawrence and Reed (2019)). However, argumentation mining has not been widely adopted in IS research, possibly due to its novel character. With my dissertation following a design science research approach, I provide transparent and artifact-based design knowledge insights into how to utilize a human-centric IS based on argumentation mining and examples for demonstration and evaluation that others can build upon.

In the subsequent sections, I will further outline how the specific design evaluations along with the empirical results contribute to the design of AI-based argumentation writing support systems and conversational argumentation learning systems. As a result of these studies, researchers and educational designers will have more insights and design knowledge to utilize adaptive argumentation learning systems in large-scale learning settings. To summarize my findings on RQ1, RQ2, RQ3 on adaptive argumentation learning systems, I will summarize the design knowledge as (1) a nascent design theory for AI-based argumentation writing support systems, and (2) a nascent design theory for conversational

argumentation learning systems according to Jones and Gregor (2007).

### 8.1.3 Components of a Design Theory for AI-based Argumentation Writing Support Systems

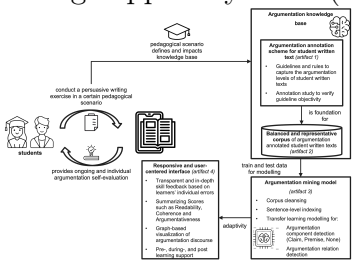
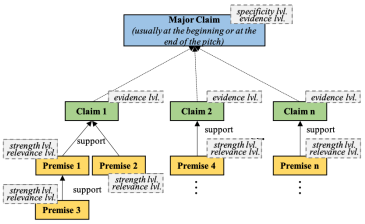
**RQ3a:** *How does an adaptive argumentation writing support system need to be designed to influence students' argumentation skills in peer review scenarios?*

To answer RQ3a, I rigorously derived requirements from theory and formulated a concise set of design principles for AI-based argumentation writing support systems. Based on 30 semi-structured user interviews, I instantiated the principles for AI-based argumentation writing support systems in two artifacts and evaluated them in qualitative pre-studies and two empirical studies. I tested the impact of my design principles for AI-based argumentation writing support systems with my instantiation ArgueLearn in a controlled laboratory experiment with 54 learners (study 1) and an extensive field experiment with 124 students over three months (study 2). I found that individuals using ArgueLearn improved their short- and long-term metacognition argumentation writing skills with better formal argumentation quality than humans who only received general or no writing support. The results indicate that learners also write texts with a higher perceived argumentation level in short-term scenarios. However, the perceived metacognitive argumentation level seems to not significantly improve through my AI-based argumentation writing support systems across different learning domains. Moreover in section 7, I evaluated my AI-based argumentation writing support system ArgumentFeedback in one qualitative study (study 1), one laboratory experiment (study 2), and one field experiment (study 3). I observed that participants who received an intelligent self-evaluation nudge including a social comparison and an adaptive feedback message when writing a persuasive business pitch wrote more formally argumentative texts compared to students who received an individual self-monitoring nudge and students who received only general self-monitoring. Furthermore, the perceived persuasiveness of the texts was significantly higher than of the texts from students using the other two tools. Moreover, the measured intention to use, the perceived ease of use, the perceived usefulness (Venkatesh and Bala, 2008), and the self-efficacy (Bandura, 1991) provided promising results for the usage of social nudging in combination with adaptive feedback for effective self-regulated argumentative skill learning.

I provide design knowledge to the scientific knowledge base in addition to the software artifact as a placed implementation of an AI-based argumentation writing support system (Matook and Brown, 2017). Design artifacts and design theories have been the most

frequent types of design science research contributions, according to Gregor and Hevner (2013). In design science, research has underlined the importance of coexisting artifact and theory contributions, but it has also revealed fundamental obstacles to accomplishing both (Gregor and Hevner, 2013; Jones and Gregor, 2007; Hevner, 2007). Research has highlighted the value of coexisting artifact and theory contributions in design science but has also identified major challenges to achieving both (Gregor and Hevner, 2013; Jones and Gregor, 2007; Hevner, 2007). The technical and scientific contributions of a design science project must be balanced (Baskerville et al., 2018). In this work, I aimed to accomplish both contribution types: I initially developed a set of design principles based on self-regulated learning theory and a technology-mediated learning perspective, which I subsequently instantiated into the technical architecture and system prototype ArgueLearn for demonstration and evaluation. (Jones and Gregor, 2007). The design principles as design theories provide functional explanations for generalized solution components of a certain IS class based on related generalized needs (Jones and Gregor, 2007). I document my design knowledge according to the eight components of a design theory by Jones and Gregor (2007) (see Table 53) to communicate my ideas to the scientific knowledge base and to capture my results. In this sense, I synthesize my theoretical contributions in the form of a "design and action" theory based on my meticulously completed design process. The goal of the theory is to establish prescriptions for form and function principles in artifact design (Jones and Gregor, 2007).

| **Component** | **Component examples** |
|---|---|
| 1) Purpose and scope | The purpose of this design theory is to formulate design principles for AI-based argumentation writing support systems to enable individuals to train their argumentation skills irrespective of instructor, time, and place through AI-based argumentation self-evaluation. The design principles are developed for educational designers. |
| 2) Constructs | An AI-based argumentation writing support system based on certain design principles of form and function increases the formal argumentation skills of learners for short- and long-term usage according to Weinberger and Fischer's (2006) scheme. Moreover, the design principles increase the technology adoption of individuals, according to Venkatesh and Bala (2008). |

| Component | Component examples |
|---|---|
| 3) Principle of form and function | Part 1 — Design principles for AI-based argumentation writing support systems:<br>• Transparent and in-depth skill feedback based on learners' individual errors (DP1).<br>• Choice of different granularity levels, visualizations, and representation of the feedback (DP2).<br>• Machine learning model and a balanced and representable corpus retrieved from a related text domain of the pedagogical scenario (DP3).<br>• Explanation about content and goals prior, during, and after the learning activity, and convey a goal and purpose of learning (DP4).<br>Part 2 — Service-oriented architecture for AI-based argumentation writing support systems (see Figure 32 in section 5):<br>  <br>Part 3 — Exemplary annotation schemes and corpora for AI-based argumentation writing support systems (see Figure 59 in section 7):<br>  |

| Component | Component examples |
|---|---|
| 4) Artifact mutability | This dissertation demonstrates how AI-based argumentation writing support systems can be instantiated in different educational scenarios in higher education. Moreover, this work evaluates how AI-based argumentation writing support systems influence argumentation skills in different experimental settings (laboratory and field settings).<br>Suggestions for improvement or adaptation of the design principles:<br>• Core design principles, such as the machine learning model and the corpora (DP3), might need to be adapted to different pedagogical scenarios (e.g., content and language of texts).<br>• Individual design principles should be evaluated in separate studies to better investigate the individual effects.<br>• Log data of AI-based argumentation writing support systems could be captured to better monitor and improve the students' argumentation learning process. |
| 5) Testable propositions | This theory provides testable propositions about the influence of AI-based argumentation writing support systems on students' argumentation skills that were tested in different empirical studies:<br>• Using an AI-based argumentation writing support system in a short-term intervention increases the formal and perceived argumentation quality of students' texts.<br>• Using an AI-based argumentation writing support system in repeated settings helps students to increase their formal metacognitive argumentation skills.<br>• Using an AI-based argumentation writing support system reduces the required amount of needed argumentation feedback from educators. |

| Component | Component examples |
|---|---|
| 6) Justificatory knowledge | AI-based argumentation writing support systems are able to improve students' formal argumentation skills based on the Toulmin argumentation model (Toulmin, 2003). Moreover, the technology-mediated learning perspective (Gupta and Bostrom, 2009) with a socio-technical system design (Bostrom and Heinen, 1977) improves students ability to self-monitor and self-evaluate their argumentation skills in different settings based on self-regulated learning theory (Bandura, 1977, 1991; Zimmerman and Schunk, 2001). |
| 7) Principles of implementation | The argumentation mining algorithm needs to be related to the domain (e.g., the task, the content, and the language) of the texts. |
| 8) Expository instantiation | The AI-based argumentation writing support system's design principles and the service-oriented architecture were instantiated in several natively built prototypes. The instantiated prototypes were evaluated in several empirical studies in controlled laboratory experiments and real-world field experiments in large-scale learning settings. |

Table 53: Design theory for AI-based argumentation writing support systems to enable individuals to train their argumentation skills irrespective of instructor, time, and place according to Jones and Gregor (2007).
Source: Own illustration

### 8.1.4 Components of a Design Theory for Conversational Argumentation Learning Systems

Besides investigating the design and the development of AI-based argumentation writing support systems, in section 5, I derived, implemented, and demonstrated design principles for the design of conversational argumentation learning systems. Thereby, I introduce the novel class of pedagogical conversational agents for argumentation skill learning. As explained in section 2, the class of pedagogical conversational agents adds a completely new perspective and paradigm to past argumentation skill learning research: dialogue-based argumentation learning. Similar to the class of AI-based argumentation writing support systems, conversational argumentation learning systems come with the characteristics of accessibility, intelligence, and socio-technical and user-centered design but also add a conversational learning paradigm to the pedagogical scenario. Hence, my findings for AI-based argumentation writing support systems are also partly true for conversational argumentation learning systems. However, the conversational interaction adds a com-

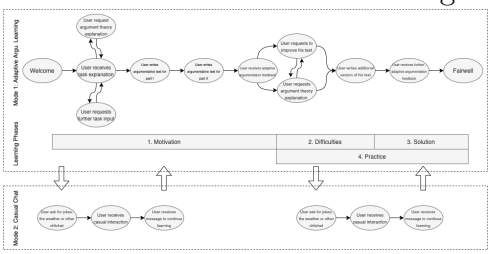
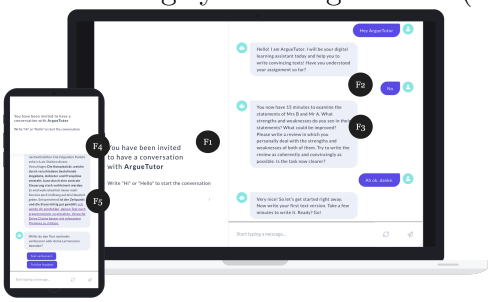
pletely new interaction layer, which significantly influences the argumentation learning process of students and thus requires further investigation (RQ3b):

**RQ3b:** *What requirements should be considered when designing an adaptive conversational argumentation learning system that aims to support students in writing argumentative texts?*

To answer RQ3b, I investigated and evaluated user-centered and theory-based design principles for conversational argumentation learning systems through the instantiation ArgueTutor following a naturalistic ex-post evaluation by Venable et al. (2016) in two empirical studies. In study 1, I compare ArgueTutor to a scripting approach for argumentation learning. I evaluated my conversational argumentation learning system in comparison with an approach that provided general argumentation feedback based on argumentation theory (Toulmin, 2003), a traditional approach for supporting persuasive writing in large-scale scenarios (e.g., Fischer et al. (2013); Scheuer et al. (2010)). I found that participants who used ArgueTutor for a persuasive writing task wrote formally more argumentative texts than the control group in this study involving 55 students. Furthermore, the perceived persuasiveness of these writings was significantly higher than that of texts created using the traditional tool. Using key constructs, I also assessed both instruments' perceived ease of use and level of enjoyment (Venkatesh et al., 2003; Venkatesh and Bala, 2008). Both constructs show great promise for using ArgueTutor as a common learning tool in lectures, according to my findings. The findings imply that ArgueTutor encourages students to produce more structured writings and inspires them to create more persuasive texts in peer learning situations like peer reviews

In study 2, I deployed two different versions of ArgueTutor in an experiment where 45 students had to compose a convincing text and were tutored by ArgueTutor (one offering individual argumentation support, the other providing non-adaptive argumentation support). The findings show that the usefulness of a conversational argumentation learning system for students comes not only in the dialogue-based interface and step-by-step coaching but also in the adaptability of the pedagogical conversational agents (see section 6).

To better structure and communicate the specific design knowledge for conversational argumentation learning systems to the scientific knowledge base, I also summarized my design findings according to the six core components of a design theory by Jones and Gregor (2007) (Table 54). Similar to the design theory for AI-based argumentation writing support systems, the theory's purpose is to provide prescriptions for principles of form and function for the artifact construction of conversational argumentation learning systems.

| Component | Component examples |
|---|---|
| 1) Purpose and scope | The objective of this design theory is to formulate design principles for AI-based conversational argumentation learning systems to support students to learn how to argue by providing individual support, guidance, and tutoring when solving argumentative writing tasks. |
| 2) Constructs | A conversational argumentation learning system based on my design principles leads to better formal and perceived argumentation skills of students for short-term usage (measured by the scheme of Weinberger and Fischer's (2006)). |
| 3) Principle of form and function | Part 1 — Design principles for conversational argu. learning systems:<br>• Pedagogical conversational agent with guidance and support for argumentative writing (DP1).<br>• Web-based pedagogical conversational agent with responsive, simple, and functional UX (DP2).<br>• Proactive argumentative tutoring and guidance with explanations based on argumentation theory (DP3).<br>• Analysis of individual argumentative components and individual feedback message (DP4).<br>• Casual chat function, e.g., with jokes or fun facts (DP5).<br>Part 2 — Interactions and dialog flow (see Figure 54 in section 6):<br>  <br>Part 3 — Prototypical instantiation of the conversational argumentation learning system ArgueTutor (see Figure 53 in section 6):<br>  |

| Component | Component examples |
|---|---|
| 4) Artifact mutability | This dissertation demonstrates how conversational argumentation learning systems can be instantiated in different educational scenarios in higher education. Moreover, this work evaluates how conversational argumentation learning systems influence argumentation skills in different experimental settings.<br>Suggestions for improvement or adaptation of the design principles:<br>• Core design principles, such as the machine learning model and the corpora (DP3), might need to be adapted to different pedagogical scenarios (e.g., content and language of texts).<br>• Human-like design elements of the pedagogical conversational agent need to be adapted based on culture or domain.<br>• Interaction dialog might change depending on culture or domain. |
| 5) Testable propositions | This theory provides testable propositions about the influence of AI-based argumentation writing support systems on students' argumentation skills that were tested in different empirical studies:<br>• Using conversational argumentation learning systems increases the student's ability to write persuasive texts independent of a human instructor.<br>• Using conversational argumentation learning systems improves the provision of guidance to students.<br>• Using conversational argumentation learning systems reduces the required amount of tutoring to write argumentative texts from human instructor. |
| 6) Justificatory knowledge | Conversational argumentation learning systems are able to trigger interactive learning behavior of students and thus lead to higher formal and perceived argumentation skills based on the Toulmin model (Toulmin, 2003). Based on a constructivist learning perspective and according to the ICAP-framework by Chi and Wylie (2014), students' interactive learning behavior leads to higher levels of learning success compared to constructive, active, and passive learning behavior. |

| Component | Component examples |
|---|---|

Table 54: Design theory for AI-based conversational argumentation learning systems according to Jones and Gregor (2007).
Source: Own illustration

Following systematic design procedure in this dissertation (Hevner et al., 2004; Peffers et al., 2007), I aimed to generate a satisfying design contribution (Gregory and Muntermann, 2014). The knowledge gained is applicable not just to the specific circumstances, but also to other scenarios such as dialogue-based reasoning and metacognition skill learning. The architectural and design concepts, for example, can be used in courses that deal with topics other than business models or languages. The back-end algorithm just needs to be changed to the new scenario. Multiple corpora and argumentation mining models, as stated, can be readily embedded as a module in the service-oriented architecture for AI-based argumentation writing support systems (for example, for English students' essays (Stab and Gurevych, 2017) or English law cases (Palau and Moens, 2009)). These use cases do not necessitate any changes to the form and function design principles or the overall system design. It's also possible to use design knowledge to teaching settings that focus on different metacognition skills. A comparable AI-based argumentation learning system can be utilized for general feedback skills or empathy abilities, for example. However, in this instance, the system architecture may need to be revised in part (e.g., the graph visualization or argumentation dashboard may need to be adapted). The research not only provides a Level 1 design science contribution by illustrating a situated artifact implementation, but it also gives a nascent design theory due to the transferability of the modular and artifact-based design knowledge (Level 2 contribution) (Gregor and Hevner, 2013). In addition to adding to the scientific knowledge base, the findings are useful for the practical use of AI techniques such as natural language processing and machine learning-based self-evaluation in education. The findings suggest that cutting-edge NLP and machine learning techniques are well-suited to sophisticated adaptive skill learning systems that can support individual learners. In contrast to traditional e-learning methodologies, I have demonstrated that advanced assistance capabilities in skill education may be implemented utilizing NLP and ML techniques.

My findings also contribute to self-regulated learning theory (Bandura, 1977, 1991; Zimmerman and Schunk, 2001). The results confirm the underlying mechanism rooted in social cognitive theory (Bandura, 1977) whereby intelligent self-evaluation combined with goal setting and formative feedback leads to positive behavior changes in learning processes. This result is especially novel for argumentation skill learning. Past research on argumentation learning systems has mostly focused on single design elements for adaptive argumentation monitoring, e.g., based on in-text highlighting (Afrin et al., 2021;

Chernodub et al., 2019; Scheuer et al., 2010; Wang et al., 2020), while perspectives on self-regulated learning theory with several design elements and self-evaluation are absent. Moreover, traditional argumentation learning theory, such as representational guidance theory (Fischer et al., 2013; Suthers, 2003), focuses exclusively on argumentation monitoring (e.g., supporting argumentation learning by providing representations of argumentation structures with the objective of stimulating and improving individual reasoning, collaboration, and learning) (Scheuer et al., 2010, 2012b; Suthers and Hundhausen, 2001). My research provides novel insights into how learners' argumentative writing based on self-evaluation may be supported in combination with a socio-technical design.

Moreover, my findings on conversational argumentation learning systems design contribute to the ICAP framework (Chi and Wylie, 2014). My studies in section 6confirm that conversational argumentation learning systems are able to trigger interactive learning behavior of students and thus lead to higher formal and perceived argumentation skills based on the Toulmin model (Toulmin, 2003). Based on the ICAP-framework by Chi and Wylie (2014), students' interactive learning behavior leads to higher levels of learning success compared to constructive, active, and passive learning behavior. My results based on two empirical studies indicate that this is also true for technology-mediated interactive learning for metacognition skills, such as argumentation.

### 8.1.5 Nudging for Self-Regulated Argumentation Learning

The last research question was concerned with the influence of digital nudging in students' argumentation self-evaluation:

**RQ3c:** *How can an argumentation writing support system be designed based on nudging and self-evaluation to influence students' argumentation writing skills?*

I answered RQ3c by constructing a novel, theory-driven argumentation annotation scheme for student-written business model pitches, compiling a corpus of 200 student-written texts, training a state of the art model to predict the individual argumentation skill level, and, finally, embedding the model in a student-centered and theory-informed design based on nudging theory. I evaluated the AI-based argumentation writing support systems ArgumentFeedback, in three studies and observed that participants who received an intelligent self-evaluation nudge including social comparison and an adaptive feedback message when writing a persuasive business pitch wrote more formally argumentative texts compared to students who received an individual self-monitoring nudge and students who received only general self-monitoring. The measured intention to use, the perceived ease of use,

the perceived usefulness (Venkatesh and Bala, 2008), and the self-efficacy (Bandura, 1991) provide promising results for the usage of social nudging in combination with adaptive feedback for effective self-regulated argumentative skill learning. The results suggest that a student-centered learning tool based on digital nudging combined with ML and NLP feedback helps students write more persuasive texts.

The findings from section 7 not only contribute by reproducing and transferring my design findings on AI-based argumentation writing support systems, but also seem to confirm the underlying mechanism rooted in social cognitive theory (Bandura, 1977), that intelligent self-evaluation with social comparison and adaptive monitoring leads to positive behavior changes in learning processes (Bandura, 1991). For argumentation skill learning, this result is especially novel. Past research on adaptive argumentation learning tools based on AI, e.g., (Afrin et al., 2021; Wang et al., 2020), have mostly focused on adaptive argumentation monitoring, e.g, based on in-text highlighting (Chernodub et al., 2019; Lippi and Torroni, 2016b; Afrin et al., 2021) or argumentation discourse monitoring. Social comparisons for argumentative writing were mostly neglected. Moreover, traditional argumentation learning theory, such as the representational guidance theory by Suthers (2003), solely focuses on argumentation monitoring, e.g., supporting argumentation learning by providing representations of argumentation structures with the objective to stimulate and improve individual reasoning, collaboration, and learning (Pinkwart et al., 2009; Suthers and Hundhausen, 2001). My research provides novel insights on how to support students' argumentative writing based on self-evaluation and digital nudging more efficiently.

Finally, my dissertation makes several contributions, especially to human—computer interaction research and educational technology. I provide insights on a holistic design of a student-centered writing support tool based on recent advantages in computational linguistics and digital nudging theory for self-regulated argumentation skill learning and metacognition skills. Moreover, I demonstrate the effectiveness of social nudging for argumentation learning through rigorously evaluating the tool ArgumentFeedback in three empirical studies with a total of 108 students. The results provide insight into the potential of designing student-centered learning tools based on NLP and ML combined with digital nudging to foster self-regulated argumentative writing in a student's learning journey. My research shows several case examples of supporting metacognition skills in a scalable and individual way in possible large-scale scenarios. Thus, I provide design knowledge for other researchers and educators to design and compare similar tools for the support of metacognition skills of students based on computational methods and nudging theory.

## 8.2 Practical Contributions of the Dissertation

In addition to its theoretical contributions, this dissertation has a practical impact on scholars and practitioners in the fields of IS, human—computer interaction, NLP, and educational technology. The practical contributions of my work are closely related to the before-mentioned theoretical contributions, especially with (1) the software instantiations, (2) the annotation schemes and corpora, and (3) the design principles alongside the findings for educators and teachers.

### 8.2.1 Software Instantiations

In this dissertation project, I designed, demonstrated, and evaluated three AI-based argumentation learning systems that provide learners with individual self-evaluation on the argumentation structure of a text irrespective of instructor, time, and place. First, I built ArgueLearn, an instantiation of an AI-based argumentation writing support system according to my theory-driven and user-based design principles (see section 5). ArgueLearn provides learners with formative argumentation feedback including argumentation text-highlighting, error highlighting and a graph-based argumentation discourse overview based on a written text. Second, I built ArgueTutor, an instantiation of a conversational argumentation learning system according to my design principles (see section 6). ArgueTutor tutors learners with adaptive and instant feedback, theoretical input, and step-by-step guidance during their writing process. Third, I built ArgumentFeedback, an instantiation of an AI-based argumentation writing support system (see section 7). ArgumentFeedback provides students with individual argumentation self-evaluation by highlighting the argumentative components, by providing an adaptive learning dashboard when writing persuasive business pitches, and by adaptively nudging students to write more argumentative texts based on social comparison (Mirsch et al., 2017; Cialdini and Goldstein, 2004). The three software instantiations contribute with their design, architect, interaction concept, and finally their real-world embedding to practitioners in the educational technology field. Educational designers can follow the same architecture or even leverage the publicly available instantiations for their own innovation projects or educational scenarios.

### 8.2.2 Annotation Schemes and Argumentation Annotated Corpora

Not only does the user interface represents a practical contribution, but also the novel annotation schemes, the ML algorithms for student-written texts as well as the publicly available corpora represent a valuable resource for researcher and designer to build

and instantiate their own AI-based argumentation writing support system, conversational argumentation learning systems or potentially new classes of argumentation learning systems.[19]

In section 5, I presented an annotated corpus of persuasive student-written peer reviews collected from a real-world learning scenario, as well as an argumentation annotation scheme. This corpus was used to create AI-powered argumentation writing support tools. Argument annotation techniques and argumentation corpora that had previously been provided had a number of flaws. First, they were rarely drawn from a contemporary learning situation; second, they lacked a systematic technique based on extensive inter-rater agreement investigations; and third, they lacked annotations of argumentation relations (see, e.g., Lawrence and Reed (2019)). With my corpus presented in section 5, I provide 1,000 peer reviews produced by students that are annotated for argument components and relationships. As a result, based on argumentation theory and earlier work on annotation schemes for convincing student essays, I contribute to NLP practitioners by deriving an annotation scheme for a new data domain for argumentation mining (Stab and Gurevych, 2014b, 2017). In addition, I offer an annotation study based on 50 compelling peer reviews annotated by three native German speakers, proving that annotation of student-generated peer reviews can be conducted reliably. Finally, I provide a publicly available corpus of 1,000 student peer reviews gathered during a German-language lecture on business innovation.

Next, in section 7, I proposed another argumentation annotation scheme and introduced a new annotated corpus of persuasive student-written business model pitches extracted from a pedagogical scenario. The final corpus consists of 200 student-written business model pitches with 3,207 sentences annotated for argument components, their relations, and six persuasiveness scores on different levels. By presenting an annotation study based on 50 persuasive pitches, I demonstrate that the annotation of student-written business model pitches is reliably possible. Both corpora, provide a valuable resource for NLP-based educational tools in research and practice. Moreover, the evaluated annotation schemes, offer researchers and practitioners valuable insight in how to model argumentative discourse in student-written texts. In fact, with the continuous rise of technology-mediated learning and especially AI for assessment, feedback, and scaffolding of students, my corpora provide a solid foundation for further inventions in the field of argumentation learning.

[19]E.g., the Argumentation Annotated Student Peer Reviews Corpus version 1.0 from section 5 is publicly available through https://github.com/thiemowa/argumentative_student_peer_reviews

### 8.2.3 Findings for Educators

Finally, the design principles, as well as the design theories, equip educators with new insights and knowledge on how to embed argumentation learning systems in different pedagogical scenarios. By framing my research around the technology-mediated learning framework by Gupta and Bostrom (2009), educators and educational institutions can build on my findings to improve the argumentation learning process based on my educational method structures investigated in this dissertation. The empirical evaluations of the instantiations of my design principles for AI-based argumentation writing support systems and conversational argumentation learning systems provide educators with profound evidence that these types of technology-mediated learning tools increase skill learning outcomes and user perception. Also, my qualitative findings provide a basis to better understand the argumentation skill learning process of students as well as the different technology-mediated methods to support and structure this process. Hence, also educational institutions can rely on quantitative and qualitative results on how to improve their distance learning scenarios.

Both, AI-based argumentation writing support systems and conversational argumentation learning systems, have a great potential to not only improve the teaching methods of educators but mostly to improve metacognition skill learning of students. Therefore, the biggest impact of my research is (hopefully) on the individual learners themselves. Students, no matter if from high school, higher education, vocational training, or continuous training, can rely on the user-centered and theory-based design of novel AI-based argumentation writing support systems and conversational argumentation learning systems, which support and tutor them in their individual learning process.

## 8.3 Limitations and Directions for Future Research

In this section, I offer areas and directions for further research. While the individual limits of the various studies, as well as closely related future research opportunities, have previously been covered, I will now focus on the overall constraints of my work on argumentation support systems, as well as research issues for future study.

I organize my suggestions for future research by the following categories: (1) research on novel pedagogical scenarios for argumentation learning systems, (2) research on the learner-centered design of argumentation learning systems, (3) research on recent technological advances for argumentation learning systems, and (4) research on ethics of AI-based learning systems.

The findings of my work are limited by several aspects. First of all, the empirical evaluation on the impact of AI-based argumentation writing support systems and conversational argumentation learning tools are only shown in two very specific pedagogical scenarios: (1) in the domain of student peer reviews on business models in German language and (2) in the domain of student business pitches in German. Further studies are needed to replicate the empirical findings on the effect of the design of these systems in other pedagogical domains (e.g., writing of persuasive discussion openings or reflective journals), in other languages (e.g., in French or Chinese), or with different design variations (e.g., empirical investigation of ArgueTutor with a casual chat mode and without a casual chat mode). The research findings are limited to the specific domains I evaluated my argumentation learning tools in. Nevertheless, even though I hypothesize that the service-orientated architecture and the design findings are also true for other pedagogical scenarios, I cannot prove it with my research design. Further investigations are needed to replicate the findings. Since my studies are also limited to long-term effects, it will be especially important to focus on longitudinal studies to investigate the continuous impact of argumentation writing support systems on learning outcomes.

Another major limitation of my results evokes from the measurements of the argumentation learning outcomes. Since metacognition skills are complex pedagogical constructs, I followed a proven method to measure the argumentation skill level of student texts by the annotation scheme of Weinberger and Fischer (2006). This annotation scheme has been widely used to measure the argumentation quality of student texts, however, is limited when it comes to capturing more sophisticated dimensions of metacognition skill learning in different domains. To more precisely measure the impact of individual argumentation feedback and self-evaluation on students' skill levels, more advanced argumentative skill assessments are necessary. These assessments should be able to capture the argumentative competencies of students in different pedagogical domains (e.g., business, ethics, and education) before and after a certain treatment on different argumentation skill dimensions (e.g. logical argumentation competencies, discourse level, and strength and relevance of argumentation structures) to represent students' argumentation skill with a higher granularity compared to the assessment method of Weinberger and Fischer (2006). Future research is needed to construct, validate, and evaluate such test measurements to more precisely measure the impact of different pedagogical learning method structures on metacognition skill learning outcomes (Gupta and Bostrom, 2009).

### 8.3.1 Research on Novel Pedagogical Scenarios for Argumentation Learning Systems

I suggest future research on argumentation learning systems in different educational contexts and didactical scenarios. As Scheuer et al. (2010) already pointed out, argumentation learning systems can be embedded in various pedagogical scenarios. Of interest would be to further investigate the embedding of argumentation support tools in collaborative learning or work scenarios. Diana (2018) investigated the impact of argumentative games to promote productive civil discourse. Fischer et al. (2013) embedded their argumentation scripting approach to promote more logical argumentation in online discussion forums. The new class of AI-based argumentation writing support systems or conversational argumentation learning systems has not been embedded in collaborative learning or work scenarios. Potential research questions arising from this dissertation include:

- How can AI-based argumentation writing support systems and conversational argumentation learning systems be embedded in collaborative learning settings?
- How can AI-based argumentation writing support systems and conversational argumentation learning systems be designed to promote civil discourse in online discussions or in social media settings?
- How can AI-based argumentation writing support systems and conversational argumentation learning systems be designed to counteract fake news and hate speech?
- What are possible use cases for AI-based argumentation writing support systems and conversational argumentation learning systems in collaborative learning settings?
- What impact have AI-based argumentation writing support systems and conversational argumentation learning systems on students' argumentation skills in collaborative settings?
- How do students perceive the interaction of AI-based argumentation writing support systems and conversational argumentation learning systems in collaborative settings?
- How should AI-based argumentation writing support systems and conversational argumentation learning systems be designed to promote long-term argumentation skill learning in collaborative settings?

The OECD called for future research on how to support metacognition skill learning (Oecd, 2018). In my dissertation, I investigate how to design novel learning tools to

support argumentation skill learning. Other metacognition skills encompass problem-solving, communication, verbal reasoning, or empathy (Oecd, 2019). My dissertation has shown that with recent technological advances, we are able to design new pedagogical scenarios to train short- and long-term metacognition skills. Hence, I suggest future research to investigate the impact of already existing NLP and ML approaches to provide students feedback and support with writing support systems or pedagogical conversational agents, for example, by leveraging empathy detection to provide students feedback on their empathy skills (Santos et al., 2018). Thus, I suggest the following questions for future research on metacognition skill training based on writing support systems and pedagogical conversational agents.

- How to design AI-based writing support systems and pedagogical conversational agents to help students to train their empathy skills?
- How to design AI-based writing support systems and pedagogical conversational agents to help students to train their creativity skills?
- How to design AI-based writing support systems and pedagogical conversational agents to help students to train their communication skills?
- How to design AI-based writing support systems and pedagogical conversational agents to help students to train their critical thinking skills?
- How should AI-based writing support systems and pedagogical conversational agents be embedded in reflective journals to promote metacognition skill learning?
- What are other didactical scenarios where AI-based writing support systems and pedagogical conversational agents could be embedded to support students to learn metacognition skills?

The learning process of argumentation can play a threefold role in certain pedagogical scenarios. First, a pedagogical scenario can aim to particular train argumentation skills, by asking learners to participate in argumentative writing activities with the pedagogical focus to promote argumentation skill learning, which was the objective of this dissertation. Second, practicing active debates and argumentation can be used as a didactical method to achieve other learning outcomes, e.g., to manifest certain factual knowledge (Jonassen and Kim, 2010; Asterhan and Schwarz, 2016). Third, actively engaging in argumentation has been shown to positively impact other metacognition skills such as communication and problem-solving in general (Duschl and Osborne, 2002; Kuhn, 2005; Van Eemeren and Grootendorst, 2016). Future research should focus on the effect of "arguing to learn"

through AI-based argumentation writing support systems and conversational argumentation learning systems. A possible scenario would be to embed AI-based argumentation writing support systems in language learning assignments to promote English language learning through argumentation feedback. Other scenarios would be to embed argumentative assessment through an AI-based argumentation writing support system to promote factual knowledge learning:

- How to design AI-based argumentation writing support systems and conversational argumentation learning systems to support English language learning of students?argumentation writing support systems
- What impact do AI-based argumentation writing support systems and conversational argumentation learning systems have on other learning outcomes?
- How to design AI-based argumentation writing support systems and conversational argumentation learning systems to factual knowledge learning through interactive exercises?
- What is the effect of argumentation feedback on factual knowledge learning?
- What is the effect of argumentation feedback on metacognition skill learning?

Last but not least, argumentation assessment is not only a prevailing need of students, but also of educators. In my dissertation, I focused on how to support students to learn how to argue. However, persuasive writing assessment of educators and essay scoring is a time-consuming task in which argumentation usually plays a crucial role (e.g., Gema et al. (2017). Future research should also focus on how to support teachers in better assessing argumentative writing excesses of their students:

- What characteristics and dimensions need to be considered when assessing students' metacognitive argumentation skills?
- How to design an assessment for metacognitive argumentation skills of students across pedagogical domains?
- How can a learning process for argumentation learning systems be conceptualized?
- How do argumentation skills form and evolve throughout a student's learning process in argumentation learning systems?
- How to design AI-based argumentation writing support systems and conversational argumentation learning systems to support educators in assessing argumentation skill levels of students?

- How to design AI-based argumentation writing support systems and conversational argumentation learning systems to support educators in essay grading?
- How can AI-based argumentation writing support systems and conversational argumentation learning systems be designed to enable educators to monitor the argumentation skill level of students?
- How can AI-based argumentation writing support systems and conversational argumentation learning systems be designed to enable educators to monitor the learning argumentation process of students?

### 8.3.2 Research on the Learner-Centered Design of Argumentation Learning Systems

Learning systems that students enjoy using, which are intuitive and easy to interact with, promote long-term usage and learning outcomes (Soloway et al., 1994). Hence, the enhancement of the interaction of argumentation learning systems based on gamification elements (Schöbel et al., 2020), nudging (Thaler and Sunstein, 2009), or other proven human—computer interaction concepts should play a significant role in future research (Xu et al., 2021). New interaction paradigms based on virtual reality, augmented reality, multi-sensory environments, or speech-based conversation provide interesting space for future research not only on AI-based argumentation writing support systems and conversational argumentation learning systems but on adaptive argumentation learning systems in general (Xu et al., 2021; Rahwan et al., 2019). Hence, I suggest devoting time to the following research questions:

- How to design argumentation learning systems based on virtual or augmented reality to help students to learn how to argue?
- How to design multi-sensor interaction systems to improve argumentation learning of students with argumentation learning systems?
- How can a speech-based conversational argumentation learning system be designed to foster students' argumentation skill learning?
- What effect have different gamification elements in AI-based argumentation writing support systems and conversational argumentation learning systems on students' perception and learning outcomes?
- What effect have different digital nudges embedded in AI-based argumentation writing support systems and conversational argumentation learning systems on students' perception and learning outcomes?

- How to measure metacognition argumentation learning outcomes more precisely?
- How to transfer the results on argumentation learning support to other pedagogical domains and languages?

### 8.3.3 Research on Recent Technological Advances for Argumentation Learning Systems

Technological advances in AI, NLP, and ML are continuing to provide new methods in how to design, build and deploy software (e.g., Alom et al. (2019); Xu et al. (2021); Rahwan et al. (2019)). Also, for argumentation learning systems recent advantages in NLP and ML bear several avenues for future research. Besides, the availability of large datasets, the vast access to computing power in the cloud or the evolution in deep learning architectures such as transformer-based models (Devlin et al., 2018) and generative adversarial networks (Creswell et al., 2018), provide enormous opportunities for enhancements of adaptive argumentation learning systems. Future research should try to leverage these novel techniques to enhance AI-based adaptive learning tools such as argumentation writing support systems and conversational argumentation learning systems in order to provide learners with an individual learning experience irrespective of an educator, time, or location. Possible research questions on technology for argumentation learning systems include the following:

- How can generative adversarial networks be embedded to provide students with more insightful argumentation suggestions or scripts?
- How can novel advances in transformer architectures be leveraged to build domain agnostic AI-based argumentation writing support systems and conversational argumentation learning systems?
- How to annotate corpora to capture domain agnostic argumentation structures?
- How to better annotate and model qualitative attributes of human argumentation?
- How can educational process mining be used to enhance and improve the argumentation learning process of students' using AI-based argumentation writing support systems and conversational argumentation learning systems?
- How can general learning analytic techniques be combined with advances in deep learning to enhance and improve the argumentation learning process of students' using AI-based argumentation writing support systems and conversational argumentation learning systems?

- How can reinforcement or active learning mechanisms be embedded to improve model accuracy of AI-based argumentation writing support systems and conversational argumentation learning systems?

### 8.3.4 Research on Ethics for AI-based Learning Tools

I aim to close my dissertation by calling for future research on ethics and biases in AI-based adaptive education. NLP and ML are susceptible to bias, resulting in systematic errors that may create unfair or discriminating predictions. For example, they can lead to inaccurate outcomes for specific subgroups or may carry the implicit values of designers and educational institutions (Rahwan et al., 2019). Moreover, AI-based learning tools operate with some level of autonomy, resulting in increased opaqueness that highlights questions of accountability and transparency (Gimpel et al., 2019). Also, these tools operate on user data and may, in fact, be used to collect enormous amounts of (sensitive) data, where user privacy becomes an even more important issue (Rahwan et al., 2019).

Taken together, while AI-based learning tools, such as argumentation writing support systems or conversational argumentation learning systems, have the potential to fundamentally improve learning scenarios and learning experience outcomes, designers and educators may need to increasingly follow ethical considerations in the design of these tools to ensure the well-being of learners and fairness of the educational assessment. Especially biases in educational datasets might result in unfair assessments of minorities. The question is not if biases exist in the underlying models and data, but rather which biases are prominently represented. For argumentation mining, Spliethöver and Wachsmuth (2020) found that argumentation debate corpora contained unbalanced and biased data, e.g., in favor of male people with European-American names. Hence, I urge future research to include ethical considerations and investigations on biases as part of their research. Following the OECD principles on AI[20], I suggest the following future research questions for AI-based educational research:

- Which biases exist in educational corpora?
- What influence have existing biases in educational corpora on students' perception and learning outcomes?
- How to prevent biases and discrimination of AI-based learning tools?
- How to design human-centered and fair AI-based learning tools?

[20]More information at https://www.oecd.org/going-digital/ai/principles/

- How to increase transparency and explainability of AI-based learning tools?
- How to increase traceability of assessments and predictions of AI-based learning tools?
- How to increase trust and communication with users of AI-based learning tools?
- How to ensure safety and security of users and their data when learning with AI-based tools?
- How to ensure accountability and auditability of AI-based learning tools?
- How to ensure a general ethical and value-sensitive design of AI-based learning tools?

# References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages. *ACM Transactions on Information Systems*, 26(3):1–34.

Ahmed Abbasi, Yilu Zhou, Shasha Deng, and Pengzhu Zhang. 2018. Text analytics to support sense-making in social media: A language-action perspective. *MIS Quarterly: Management Information Systems*, 42(2):427–464.

Alessandro Acquisti, Saranga Komanduri, Civis Analytics, Pedro Giovanni Leon, Banco De Mexico, Florian Schaub, Yang Wang, Shomir Wilson, ; L F Cranor, N Sadeh, ; Y Wang, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Norman Sadeh, and Manya Sleeper. 2017. 44 Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online ACM Reference Format. *ACM Computing Surveys*, 50(3).

Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA. Acm.

Basant Agarwal and Namita Mittal. 2016. Prominent Feature Extraction for Sentiment Analysis. *Prominent Feature Extraction for Sentiment Analysis*, pages 21–45.

Ritu Agarwal and Elena Karahanna. 2000. Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS Quarterly*, 24(4):665.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Maryam Alavi and Dorothy E. Leidner. 2001. Research Commentary: Technology-Mediated Learning - A Call for Greater Depth and Breadth of Research. *Information Systems Research*, 12(1):1–10.

Mehdi Allahyari, Elizabeth D Trippe, and Juan B Gutierrez. 2017. A Brief Survey of Text Mining : Classification , Clustering and Extraction Techniques.

Md Zahangir Alom, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C. Van Essen, Abdul A. S. Awwal, Vijayan K. Asari, Md Zahangir Alom, Tarek M. Taha, Chris Yakopcic, Stefan Westberg,

Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C. Van Essen, Abdul A. S. Awwal, and Vijayan K. Asari. 2019. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, 8(3):292.

Salha M Alzahrani, Naomie Salim, Ajith Abraham, and Senior Member. 2012. Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. In *Ieee Transactions On Systems, Man, And Cybernetics*, volume 42, page 133.

Anaconda. 2020. State of Data Science 2020.

John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray. Pelletier. 1995. Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4(2):167–207.

Lorin W. Anderson and Lauren A. Sosniak. 1996. Bloom's Taxonomy: A Forty-Year Retrospective. *History of Education Quarterly*, 36(1):76.

Nicholas O Andrews and Edward A Fox. 2007. Recent developments in document clustering. *Citeseer*, pages 1–25.

Hilary Arksey and Lisa O'Malley. 2005. Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, 8(1):19–32.

S. J. Ashford. 1986. Feedback-Seeking in Individual Adaptation : A Resource Perspective. *Academy of Management Journal*, 29(3):465–487.

Kevin D Ashley and Vern R Walker. 2013. Toward Constructing Evidence-based Legal Arguments Using Legal Decision Documents and Machine Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, Icail '13, pages 176–180, New York, NY, USA. Acm.

Christa S.C. Asterhan and Baruch B. Schwarz. 2009. Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. *Cognitive Science*, 33(3):374–400.

Christa S.C. Asterhan and Baruch B. Schwarz. 2016. Argumentation for Learning: Well-Trodden Paths and Unexplored Territories. *Educational Psychologist*, 51(2):164–187.

R. C. Atkinson and R. M. Shiffrin. 1968. Human Memory: A Proposed System and its Control Processes. *Psychology of Learning and Motivation - Advances in Research and Theory*, 2(C):89–195.

Roger Azevedo and Allyson F. Hadwin. 2005. Scaffolding self-regulated learning and metacognition - Implications for the design of computer-based scaffolds.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.

R. (Ricardo) Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. ACM Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

K.D: Bailey. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. Sage, Newbury Park.

Albert Bandura. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2):191–215.

Albert Bandura. 1991. Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2):248–287.

Richard Baskerville, Abayomi Baiyere, Shirley Gregor, Alan Hevner, and Matti Rossi. 2018. Design science research contributions: Finding a balance between artifact and theory. *Journal of the Association for Information Systems*, 19(5):358–376.

Konstantin Bauman and Alexander Tuzhilin. 2018. Recommending Remedial Learning Materials to Students by Filling Their Knowledge Gaps. *MIS Quarterly*, 42(1):313–332.

Bradford S. Bell and Steve W.J. Kozlowski. 2008. Active Learning: Effects of Core Training Design Elements on Self-Regulatory Processes, Learning, and Adaptability. *Journal of Applied Psychology*, 93(2):296–316.

Philip Bell. 1997. Using Argument Representations to Make Thinking Visible for Individuals and Groups. In *Proceedings of the 2nd International Conference on Computer Support for Collaborative Learning*, Cscl '97, page 10–19. International Society of the Learning Sciences.

Izak Benbasat and Robert W. Zmud. 2003. The identity crisis within the IS discipline: Defining and communicating the discipline's core properties.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*.

Or Biran and Owen Rambow. 2011. Identifying Justifications in Written Dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, Icsc '11, pages 162–168, Washington, DC, USA. IEEE Computer Society.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, volume 43.

Robert A. Bjork, John Dunlosky, and Nate Kornell. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64(November 2012):417–444.

Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1):5–31.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: a Corpus of Non-Native English. *ETS Research Report Series*, 2013(2):i–15.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management.

Robert P. Bostrom and J. Stephen Heinen. 1977. MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly*, 1(3):17.

David Brett and Antonio Pinna. 2015. Patterns, Fixedness and Variability: Using PoS-grams to Find Phraseologies in the Language of Travel Journalism. *Procedia - Social and Behavioral Sciences*, 198(Cilc):52–57.

Jan Brocke, Alexander Simons, Kai Riemer, and Ralf Plattfaut. 2015. Standing on the Shoulders of Giants : Challenges and Recommendations of Literature Search in Information Systems Research Standing on the Shoulders of Giants : Challenges and Recommendations. 37.

Kenneth G. Brown, Garett Howardson, and Sandra L. Fisher. 2016. Learner Control and e-Learning: Taking Stock and Moving Forward. *Annual Review of Organizational Psychology and Organizational Behavior*, 3:267–291.

Ruth M. J. Byrne. 1989. Everyday Reasoning with Conditional Sequences. *The Quarterly Journal of Experimental Psychology Section A*, 41(1):141–166.

Elena Cabrio and Serena Villata. 2012. Natural Language Arguments: A Combined Approach. In *20th European Conference on Artificial Intelligence (ECAI 2012)*, Montpellier, France.

Elena Cabrio and Serena Villata. 2014. Towards a Benchmark of Natural Language Arguments. *CoRR*, abs/1405.0.

William Cai, Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian, Zheng Wei, Joseph Jay Williams, and Sharad Goel. 2019. MathBot: A Personalized Conversational Agent for Learning Math.

Cansu Canca. 2020. Operationalizing Ai Ethics Principles ∗. *Communications of the ACM*, 63(12):3–6.

Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A review of technology-mediated nudging in human-computer interaction. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–15.

Ana Caraban, Loukas Konstantinou, and Evangelos Karapanos. 2020. The nudge deck: A design support tool for technology-mediated nudging. In *DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 395–406.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:621–631.

Thomas F. Carolan, Shaun D. Hutchins, Christopher D. Wickens, and John M. Cumming. 2014. Costs and benefits of more learner freedom: Meta-analyses of exploratory and learner control training methods. *Human Factors*, 56(5):999–1004.

Leona Chandra, Stefan Seidel, and Shirley Gregor. 2015. Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2015-March:4039–4048.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. 2000. Crisp-dm 1.0. *CRISP-DM consortium*.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural Argument Mining at Your Fingertips. pages 195–200.

Michelene T.H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49(4):219–243.

Glenn Rowe Chris Reed Raquel Mochales Palau and Marie-Francine Moens. 2008. Language Resources for Studying Argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Robert B. Cialdini and Noah J. Goldstein. 2004. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*.

Mike Cohn. 2004. User Stories Applied For Agile Software Development. Technical report.

Harris M. Cooper. 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, 1(1):104–126.

Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Jean Valere Cossu, Nicolas Dugue, and Vincent Labatut. 2015. Detecting Real-World Influence through Twitter. *Proceedings - 2nd European Network Intelligence Conference, ENIC 2015*, pages 83–90.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. 2018. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, 35(1):53–65.

Peter Daly and Dennis Davy. 2016. Structural, linguistic and rhetorical features of the entrepreneurial pitch: Lessons from Dragons' Den. *Journal of Management Development*, 35(1):120–132.

Ben Daniel. 2015. Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, 46(5):904–920.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim ? Cross-Domain Claim Identification. *Emnlp*.

R De Groot, R Drachman, R Hever, B Schwartz, U Hoppe, A Harrer, M De Laat, R Wegerif, B M Mclaren, and B Baurens. 2007. Computer Supported Moderation of E-Discussions: the ARGUNAUT Approach. Technical report.

Gerardine DeSanctis and Marshall Scott Poole. 1994. Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory. *Organization Science*, 5(2):121–147.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Nicholas Diana. 2018. Leveraging educational technology to improve the quality of civil discourse. In *International Conference on Artificial Intelligence in Education*, volume 10948 Lnai, pages 517–520. Springer Verlag.

Pierre Dillenbourg. 2002a. Over-scripting CSCL: The risks of blending collaborative learning with instructional design.

Pierre Dillenbourg. 2002b. Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In Three worlds of CSCL: Can we support CSCL. *Three worlds of CSCL. Can we support CSCL?*, (March):61–91.

Pierre Dillenbourg and Fabrice Hong. 2008. The mechanics of CSCL macro scripts. *International Journal of Computer-Supported Collaborative Learning*, 3(1):5–23.

Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. 2009. The Evolution of Research on Computer-Supported Collaborative Learning. In Nicolas Balacheff, Sten Ludvigsen, Ton de Jong, Ard Lazonder, and Sally Barnes, editors, *Technology-Enhanced Learning: Principles and Products*, pages 3–19. Springer Netherlands, Dordrecht.

Nicola Döring and Jürgen Bortz. 2015. Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. page 1051.

D. Harold Doty and William H. Glick. 1994. Typologies as a Unique Form of Theory Building: Toward Improved Understanding and Modeling. *The Academy of Management Review*, 19(2):230.

Rosalind Driver, Paul Newton, and Jonathan Osborne. 2000. Establishing the norms of scientific argumentation in classrooms. *Science Education*, 84(3):287–312.

Ana Maria Ducasse. 2020. Evidence-based persuasion: A cross-cultural analysis of entrepreneurial pitch in English and Spanish. *Journal of International Entrepreneurship*, 18(4):492–510.

Richard A. Duschl and Jonathan Osborne. 2002. Supporting and Promoting Argumentation Discourse in Science Education. *Studies in Science Education*, 38(1):39–72.

Frans H. van Eemeren, Rob Grootendorst, Ralph H. Johnson, Christian Plantin, Charles A. Willard, Rob Grootendorst, Ralph H. Johnson, Christian Plantin, and Charles A. Willard. 1996. *Fundamentals of Argumentation Theory*. Routledge.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Corpus Wide Argument Mining – a Working Solution. (1).

eMarketer. 2017. Alexa , Say What?! Voice-Enabled Speaker Usage to Grow Nearly 130% This Year. Technical report.

Peggy A. Ertmer and Timothy J. Newby. 1993. Behaviorism, Cognitivism, Constructivism: Comparing Critical Features from an Instructional Design Perspective. *Performance Improvement Quarterly*, 6(4):50–72.

Lauren Eskreis-Winkler, Katherine L. Milkman, Dena M. Gromet, and Angela L. Duckworth. 2019. A large-scale field experiment shows giving advice improves academic outcomes for the advisor. *Proceedings of the National Academy of Sciences of the United States of America*, 116(30):14808–14810.

Charles Fadel, Maya Bialik, and Bernie Trilling. 2015. *Four-dimensional education : the competencies learners need to succeed*.

Ya-ching Fan, Tzu-hua Wang, and Kuo-hua Wang. 2020. Studying the effectiveness of an online argumentation model for improving undergraduate students ' argumentation ability. (December 2019):1–14.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From Data Mining to Knowledge Discovery in Databases.

Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human Computer Studies*, 132:138–161.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. pages 276–284.

José Santiago Fernández-Vázquez and Roberto Carlos Álvarez-Delgado. 2019. The interaction between rational arguments and emotional appeals in the entrepreneurial pitch. *International Journal of Entrepreneurial Behaviour and Research*, 26(3):503–520.

Frank Fischer, Ingo Kollar, Karsten Stegmann, and Christof Wecker. 2013. Toward a Script Theory of Guidance in Computer-Supported Collaborative Learning. *Educational Psychologist*, 48(1):56–66.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jürgen Flender, Ursula Christmann, and Norbert Groeben. 1999. Entwicklung und erste Validierung einer Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 20(4):309–325.

R Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education*, 897.

Eirini Florou, Stasinos Konstantopoulos, Antonis Kukurikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54. Acl.

Forbes. 2017. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says.

James B Freeman. 2001. Argument structure and disciplinary perspective. *Argumentation*, 15(4):397–423.

Hansjörg Fromm, Thiemo Wambsganss, and Matthias Söllner. 2019. Towards a Taxonomy of Text Mining Features. In *In Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm, Sweden.

Johannes Fürnkranz. 1999. Exploiting Structural Information for Text Classification on the WWW. pages 487–497.

Aryo Pradipta Gema, Suhendro Winton, Theodorus David, Derwin Suhartono, Muhsin Shodiq, and Wikaria Gazali. 2017. It Takes Two to Tango: Modification of Siamese Long Short Term Memory Network with Attention Mechanism in Recognizing Argumentative Relations in Persuasive Essay. *Procedia Computer Science*, 116:449–459.

Henner Gimpel, Sarah Bayer, Elisabeth André, Ivo Benke, Alexander Benlian, Nicholas Cummins, Oliver Hinz, Kristian Kersting, Alexander Maedche, Jan Riemann, Björn Schuller, and Klaus Weber. 2019. Humane Anthropomorphic Agents : the Quest for the Outcome Measure. In *AIS SIGPrag 2019 pre-ICIS workshop "Values and Ethics in the Digital Age"*.

Jochen. Glaŝer and Grit. Laudel. 2010. *Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen*. VS Verlag fuï Sozialwiss.

Shirley Gregor, Leona Chandra Kruse, and Stefan Seidel. 2020. Research perspectives: The anatomy of a design principle. *Journal of the Association for Information Systems*, 21(6):1622–1652.

Shirley Gregor and Alan R Hevner. 2013. Positioning and Presenting Design Science Research for Maximum Impact.

Robert Wayne Gregory and Jan Muntermann. 2014. Research Note: Heuristic Theorizing: Proactively Generating Design Theories.

Arne Gruettner, Min Vitisvorakarn, Thiemo Wambsganss, Roman Rietsche, and Andrea Back. 2020. The new window to athletes' soul - What social media tells us about

athletes' performances. *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Arne Gruettner, Thiemo Wambsganss, and Andrea Back. 2021. From data to dollar: Using the wisdom of an online tipster community to improve sports betting returns. *European Journal of International Management*, 15(2-3):314–338.

Serkan Günal, Semih Ergin, M. Bilginer Gülmezoğlu, and Ö Nezih Gerek. 2006. On feature extraction for spam e-mail detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4105 Lncs, pages 635–642. Springer Verlag.

Robert Gunning. 1969. The Fog Index After Twenty Years. *Journal of Business Communication*, 6(2):3–13.

Saurabh Gupta and Robert P. Bostrom. 2009. Technology-Mediated learning: A comprehensive theoretical model. *Journal of the Association for Information Systems*, 10(9):686–714.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting Debate Portals for Semi-supervised Argumentation Mining in User-Generated Web Discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

Donald C. Hambrick. 1984. Taxonomic Approach to Studying Strategy: Some Conceptual and Methodological Issues. *Journal of Management*, 10(1):27–41.

K.M. Hammouda and M.S. Kamel. 2004. Efficient phrase-based document indexing for Web document clustering. *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1279–1296.

Julia Hancke, Sowmya V., and Detmar Meurers. 2012. Readability Classification for German using Lexical, Syntactic, and Morphological Features. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 1063–1080.

Katrin Hartwig and Christian Reuter. 2021. Nudging users towards better security decisions in password creation using whitebox-based multidimensional visualisations. *Behaviour and Information Technology*.

Hussein Hashimi, Alaaeldin Hafez, and Hassan Mathkour. 2015. Selection criteria for text mining approaches. *Computers in Human Behavior*, 51:729–733.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. pages 299–305.

Alan R Hevner. 2007. A three cycle view of design science research. *Scandinavian Journal of Information Systems*, pages 1–6.

Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. 2004. Design Science in Information Systems Research. *Design Science in IS Research MIS Quarterly*, 28(1):75.

Sebastian Hobert. 2019. Say hello to 'Coding Tutor'! Design and evaluation of a chatbot-based learning system supporting students to learn to program. *40th International Conference on Information Systems, ICIS 2019*, pages 1–17.

Sebastian Hobert and Raphael Meyer Von Wolff. 2019. Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. *14th International Conference on Wirtschaftsinformatik, Siegen, Germany.*

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2018. Fine-grained essay scoring of a complex writing task for native speakers. pages 357–366.

Hospice Houngbo and Robert Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 19–23, Baltimore, Maryland. Association for Computational Linguistics.

Garett N. Howardson, Karin A. Orvis, Sandra L. Fisher, and Michael E. Wasserman. 2017. The psychology of learner control in training: A multilevel, interactionist framework. In *The Cambridge Handbook of Workplace Training and Employee Development*, pages 98–120. Cambridge University Press.

Chenn Jung Huang, Shun Chih Chang, Heng Ming Chen, Jhe Hao Tseng, and Sheng Yuan Chien. 2016. A group intelligence-based asynchronous argumentation learning-assistance platform. *Interactive Learning Environments*, 24(7):1408–1427.

Ni Huang, Jiayin Zhang, Gordon Burtch, Xitong Li, and Peiyu Chen. 2021. Combating Procrastination on Massive Online Open Courses via Optimal Calls to Action. *Information Systems Research*, 32(2):301–317.

Juhani Iivari. 2015. Distinguishing and contrasting two strategies for design science research.

Kia Jahanbin and Vahid Rahmanian. 2020. Using twitter and web news mining to predict COVID-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, (March):26–28.

Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. 2019. Ladders for Learning: Is Scaffolding the Key to Teaching Problem Solving in Technology-mediated Learning Contexts? *Academy of Management Learning & Education*.

Jiwoon Jeon, W Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. A Framework to Predict the Quality of Answers with Non-Textual Features. Technical report.

Patrick Jermann, Pierre Dillenbourg, and Jean-Christophe Brouze. 1999. Dialectics for collective activities: an approach to virtual campus design. Technical report.

Ansamma John, P. S. Premjith, and M. Wilscy. 2017. Extractive multi-document summarization using population-based multicriteria optimization. *Expert Systems with Applications*, 86:385–397.

Kent Johnson. 2007. An Overview of Lexical Semantics. *Philosophy Compass*, 3(1):119–134.

Steven L. Johnson, Hani Safadi, and Samer Faraj. 2015. The emergence of online community Leadership. *Information and Organization*, (July):35–68.

David H. Jonassen and Bosung Kim. 2010. Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development*, 58(4):439–457.

David H. Jonassen and Susan M. Land. 2012. *Theoretical foundations of learning environments*. Routledge.

David Jones and Shirley Gregor. 2007. The anatomy of a design theory. *Journal of the Association for Information Systems*, 8(5):312–355.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* -, pages 22–es, Morristown, NJ, USA. Association for Computational Linguistics.

Leonard. Kaufman and Peter J. Rousseeuw. 2005. *Finding groups in data : an introduction to cluster analysis*. Wiley.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting Argument persuasiveness in student essays. *IJCAI International Joint Conference on Artificial Intelligence*, 2018-July:4130–4136.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays. pages 3994–4004.

Allen Kent, Madeline M. Berry, Fred U. Luehrs, and J. W. Perry. 1955. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2):93–101.

Alice Kerly, Phil Hall, and Susan Bull. 2007. Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, 20(2):177–185.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Chan Min Kim and Amy L. Baylor. 2008. A virtual change agent: Motivating pre-service teachers to integrate technology in their future classrooms. *Educational Technology and Society*, 11(2):309–321.

Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys effects of platform and conversational style on survey response quality. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–12.

J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*.

Roger E. Kirk. 1982. Procedures for testing hypotheses about means: Rules governing the choice of error term for simple effects means and simple main effects. In *Experimental design: Procedures for the behavioral sciences*, 2nd edition, page 509–510. Brooks/Cole, Belmont, CA.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications. In *Proceedings of the 2nd Workshop on Argumentation Mining held in conjunction with the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, pages 1–11.

Ingo Kollar, Frank Fischer, and Friedrich W. Hesse. 2006. Collaboration Scripts – A Conceptual Analysis. *Educational Psychology Review 2006 18:2*, 18(2):159–185.

Timothy Koschmann. 1996. Paradigm Shifts and Instructional Technology. Technical report.

Tobias Kowatsch, Marcia Nißen, Chen-hsuan Iris Shih, Dominik Rüegger, Andreas Filler, Florian Künzler, Filipe Barata, Severin Haug, Björn Brogle, Katrin Heldt, Pauline Gindrat, Nathalie Farpour-lambert, and Dagmar Allemand. 2017. Text-based Healthcare Chatbots Supporting Patient and Health Professional Teams : Preliminary Results of a Randomized Controlled Trial on Childhood Obesity. *Persuasive Embodied Agents for Behavior Change (PEACH2017) Workshop*, 1(Iva 2017):1–10.

Kurt Kraiger and Eddie Jerden. 2008. A meta-analytic investigation of learner control: Old findings and new directions. In *Toward a science of distributed learning.*, pages 65–90. American Psychological Association.

Aliane Loureiro Krassmann, Fábio Josende Paz, Clóvis Silveira, Liane Margarida Rockenbach Tarouco, and Magda Bercht. 2018. Conversational Agents in Distance Education: Comparing Mood States with Students' Perception. *Creative Education*, 09(11):1726–1742.

David R Krathwohl. 2002. A Revision Of Bloom's Taxonomy Of Educational Objectives. *Theory into Practice*, 41(4):302.

Klaus Krippendorf. 1980. Content analysis : an introduction to its methodology.

Klaus Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Departmental Papers (ASC)*, 38.

Deanna Kuhn. 1992. Thinking as Argument. *Harvard Educational Review*, 62(2):155–179.

Deanna Kuhn. 1993. Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, 77(3):319–337.

Deanna. Kuhn. 2005. *Education for thinking*. Harvard University Press.

James A. Kulik and J. D. Fletcher. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, 86(1):42–78.

Dennis Kundisch, Jan Muntermann, Anna Maria Oberländer, Daniel Rau, Maximilian Röglinger, Thorsten Schoormann, and Daniel Szopinski. 2021. An Update for Taxonomy Designers. *Business & Information Systems Engineering 2021*, pages 1–19.

Severin Landolt, Thiemo Wambsganß, and Matthias Söllner. 2021. A taxonomy for deep learning in natural language processing. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

Sven Laumer, Christian Maier, and Fabian Tobias Gubler. 2019. Chatbot Acceptance in Healthcare: Explaining User Adoption of Conversational Agents for Disease Diagnosis. *Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden*, pages 0–18.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2019. ArguminSci: A Tool for Analyzing Argumentation and Rhetorical Aspects in Scientific Writing. pages 22–28.

E. J. Lawrence, P. Shaw, D. Baker, S. Baron-Cohen, and Anthony S. David. 2004. Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychological Medicine*, 34(5):911–919.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Doug Lederman. 2018. New data: Online enrollments grow, and share of overall enrollment grows faster.

Matthew K.O. Lee, Christy M.K. Cheung, and Zhaohui Chen. 2005. Acceptance of Internet-based learning medium: The role of extrinsic and intrinsic motivation. *Information and Management*, 42(8):1095–1104.

Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining behavioral economics to design persuasive technology for healthy choices. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 325–334.

Katja Lehmann, Matthias Söllner, and Jan Marco Leimeister. 2016. Design and Evaluation of an IT-based Peer Assessment to Increase Learner Performance in Large-Scale Lectures. *ICIS 2016 Proceedings*.

Thomas Lehmann, Inka Hähnlein, and Dirk Ifenthaler. 2014. Cognitive, metacognitive and motivational perspectives on preflection in self-regulated online learning. *Computers in Human Behavior*, 32:313–323.

Dorothy E. Leidner and Sirkka L. Jarvenpaa. 1995. The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly: Management Information Systems*, 19(3):265–291.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. Technical report.

Shasha Liao and Ralph Grishman. 2010. Large Corpus-based Semantic Feature Extraction for Pronoun Coreference. Technical report.

Nedim Lipka and Benno Stein. 2010. Identifying featured articles in Wikipedia: Writing style matters. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 1147–1148, New York, New York, USA. ACM Press.

Marco Lippi and Paolo Torroni. 2015a. Argument mining: A machine learning perspective. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9524, pages 163–176. Springer, Cham.

Marco Lippi and Paolo Torroni. 2015b. Argumentation Mining: State of the Art and Emerging Trends. *IJCAI International Joint Conference on Artificial Intelligence*, 2015-Janua(2):4207–4211.

Marco Lippi and Paolo Torroni. 2016a. Argument Mining from Speech: Detecting Claims in Political Debates.

Marco Lippi and Paolo Torroni. 2016b. MARGOT: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303.

Babak Loni, Gijs Van Tulder, Pascal Wiggers, David M.J. Tax, and Marco Loog. 2011. Question classification by weighted combination of lexical, syntactic and semantic features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6836 Lnai, pages 243–250. Springer, Berlin, Heidelberg.

Rutal Mahajan and Mukesh Zaveri. 2017. SVNIT SemEval 2017 Task-6: Learning a Sense of Humor Using Supervised Approach. pages 411–415.

Sandra L. Marcus and Lance J. Rips. 1979. Conditional reasoning. *Journal of Verbal Learning and Verbal Behavior*, 18(2):199–223.

Sabine Matook and Susan A. Brown. 2017. Characteristics of IT artifacts: a systems thinking-based framework for delineating and theorizing IT artifacts. *Information Systems Journal*, 27(3):309–346.

Richard E. Mayer. 2003. The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, 13(2):125–139.

Richard E. Mayer. 2009. *Multimedia Learning*. Cambridge University Press, Cambridge.

Richard E. Mayer. 2014. Cognitive theory of multimedia learning. *The Cambridge Handbook of Multimedia Learning, Second Edition*, (May):43–71.

Simon McAlister, Andrew Ravenscroft, and Eileen Scanlon. 2004. Combining interaction and context design to support collaborative argumentation using a tool for synchronous CMC. *Journal of Computer Assisted Learning*, 20(3):194–204.

Joseph E. Mcgrath. 1995. METHODOLOGY MATTERS: DOING RESEARCH IN THE BEHAVIORAL and SOCIAL SCIENCES. In *Readings in Human–Computer Interaction*, pages 152–169. Morgan Kaufmann.

Margaret M. Mcmanus and Robert M. Aiken. 1996. Teaching collaborative skills with a group leader computer tutor. *Education and Information Technologies*, 1(1):75–96.

Albert Meijer and Martijn Wessels. 2019. Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration*, 42(12):1031–1039.

Janet Metcalfe. 2017. Learning from Errors. *Annual Review of Psychology*, 68:465–489.

Jack Mezirow. 1991. *Transformative dimensions of adult learning.* Jossey-Bass, San Francisco, CA 94104-1310.

Editors D Michie, D J Spiegelhalter, C C Taylor, and C.C. Taylor D. Michie, D. J. Spiegelhalter. 1994. Machine Learning , Neural and Statistical Classification. *Proceeding.*

Tobias Mirsch, Christiane Lehrer, and Reinhard Jung. 2017. Digital Nudging: Altering User Behavior in Digital Environments. In *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017).*

Tobias Mirsch, Christiane Lehrer, and Reinhard Jung. 2018. Making digital nudging applicable: The digital nudge design method. In *International Conference on Information Systems 2018, ICIS 2018*. Association for Information Systems. AIS Electronic Library (AISeL).

Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. 2013. Opinion mining: reviewed from word to document level. *Social Network Analysis and Mining*, 3(1):107–125.

Raquel Mochales and Marie-Francine Moens. 2008. Study on the Structure of Argumentation in Case Law. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 11–20, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-francine Moens. 2016. Argumentation Mining: Where are we now, where do we want to be and how do we get there? *ACM Transactions on Internet Technology*, 16(2):1–25.

Marie-francine Moens, Erik Boiy, and Chris Reed. 2007. Automatic Detection of Arguments in Legal Texts. *Proceedings of the 11th international conference on Artificial intelligence and law. ACM*, (January).

Mehryar. Mohri, Afshin. Rostamizadeh, and Ameet. Talwalkar. 2013. *Foundations of machine learning*. MIT Press.

Alessandro Moschitti and Roberto Basili. 2004. Complex Linguistic Features for Text Classification: A Comprehensive Study. pages 181–196. Springer, Berlin, Heidelberg.

Benjamin Motz, Matthew G. Mallon, and Joshua D. Quick. 2021. Automated Educative Nudges to Reduce Missed Assignments in College. *IEEE Transactions on Learning Technologies*.

Benjamin Mueller and Nils Urbach. 2017. Understanding the why, what, and how of theories in IS research. *Communications of the Association for Information Systems*, 41(1):349–388.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In *Coling*.

S. NagaPrasad, V. B. Narsimha, P. Vijayapal Reddy, and A. Vinaya Babu. 2015. Influence of lexical, syntactic and structural features and their combination on Authorship Attribution for Telugu Tex. In *Procedia Computer Science*, volume 48, pages 58–64. Elsevier B.V.

Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1):81–103.

Sapna Negi and Paul Buitelaar. 2014. INSIGHT Galway: Syntactic and Lexical Features for Aspect Based Sentiment Analysis. pages 346–350.

Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5892–5899.

Robert C. Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

Naghmeh Niknejad, Waidah Ismail, Imran Ghani, Behzad Nazari, Mahadi Bahari, and Ab Razak Bin Che Hussin. 2020. Understanding Service-Oriented Architecture (SOA): A systematic literature review and directions for further investigation.

Michael Noetel, Shantell Griffith, Oscar Delaney, Nicole Rose Harris, Taren Sanders, Philip Parker, Borja del Pozo Cruz, and Chris Lonsdale. 2021. Multimedia design for learning: An overview of reviews with meta-meta-analysis. *Review of Educational Research*.

Omid Noroozi, Paul A. Kirschner, Harm J.A. Biemans, and Martin Mulder. 2018. Promoting Argumentation Competence: Extending from First- to Second-Order Scaffolding Through Adaptive Fading.

Alfred North Whitehead and Bertrand Russell. 1910. *Principia Mathematica*. Cambridge University Press}.

Jay F. Nunamaker, Robert O. Briggs, Douglas C. Derrick, and Gerhard Schwabe. 2015. The Last Research Mile: Achieving Both Rigor and Relevance in Information Systems Research. *Journal of Management Information Systems*, 32(3):10–47.

E. Michael Nussbaum, Denise L. Winsor, Yvette M. Aqui, and Anne M. Poliquin. 2007. Putting the pieces together: Online argumentation vee diagrams enhance thinking during discussions. *International Journal of Computer-Supported Collaborative Learning*, 2(4):479–500.

Benjamin D. Nye. 2014. Barriers to ITS Adoption: A systematic mapping study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8474 Lncs, pages 583–590. Springer Verlag.

Stephen D. O'Connell and Guido Lang. 2018. Can Personalized Nudges Improve Learning in Hybrid Classes? Experimental Evidence From an Introductory Undergraduate Course. *Journal of Research on Technology in Education*, 50(2):105–119.

Angela M. O'Donnell, Donald F. Dansereau, and Richard H. Hall. 2002. Knowledge maps as scaffolds for cognitive processing. *Educational Psychology Review*, 14(1):71–86.

Oecd. 2018. The Future of Education and Skills - Education 2030.

Oecd. 2019. OECD learning compass 2030: In brief.

S. Oh and D.H. Jonassen. 2007. Scaffolding online argumentation during problem solving. *Journal of Computer Assisted Learning*, 23(2):95–110.

David Opitz and Richard Maclin. 1999. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligent Research*, 11:169–198.

Jonathan F. Osborne, J. Bryan Henderson, Anna MacPherson, Evan Szu, Andrew Wild, and Shi Ying Yao. 2016. The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, 53(6):821–846.

Alexander Osterwalder, Yves Pigneur, Greg Bernarda, and Alan Smith. 2014. *Value proposition design : how to create products and services customers want : get started with ...*

Robert Östling, André Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated Essay Scoring for Swedish. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.

Ibrahim Burak Ozyurt. 2012. Automatic Identification and Classification of Noun Argument Structures in Biomedical Literature. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(6):1639–1648.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Sinno Jialin Pan and Qiang Yang. 2009. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–15.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135.

Sabine Payr. 2003. The virtual university's faculty: An overview of educational agents. *Applied Artificial Intelligence*, 17(1):1–19.

Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research.

Reinhard Pekrun and Elizabeth J. Stephens. 2012. Academic emotions. *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.*, pages 3–31.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1532–1543. Association for Computational Linguistics (ACL).

Chaim Perelman. 1971. The New Rhetoric. In Yehoshua Bar-Hillel, editor, *Pragmatics of Natural Languages*, pages 145–149. Springer Netherlands, Dordrecht.

Isaac Persing and Vincent Ng. 2016. End-to-End Argumentation Mining in Student Essays. In *Naacl Hlt*, pages 1384–1394.

Vladimir Pestov. 2013. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10):1427–1437.

Jean Piaget, Terrance Brown, and Kishore Julian Thampy. 1986. The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development. Jean Piaget , Terrance Brown , Kishore Julian Thampy. *American Journal of Education*, 94(4):574–577.

Niels Pinkwart, Vincent Aleven, Kevin Ashley, and Collin Lynch. 2006. Schwachstellenermittlung und Rückmeldungsprinzipen in einem intelligenten Tutorensystem für juristische Argumentation. Technical report.

Niels Pinkwart, Kevin Ashley, Collin Lynch, and Vincent Aleven. 2009. Evaluating an Intelligent Tutoring System for Making Legal Arguments with Hypotheticals. Technical report.

Philip M. Podsakoff and Jiing Lih Farh. 1989. Effects of feedback sign and credibility on goal setting and task performance. *Organizational Behavior and Human Decision Processes*, 44(1):45–67.

John L Pollock. 1995. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA, USA.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Prakash Poudyal, Teresa Goncalves, and Paulo Quaresma. 2016. Experiments On Identification of Argumentative Sentences. *10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, pages 398–403.

James Pustejovsky and Amber Stubbs. 2013. *Natural Language Annotation*. O'Reilly.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, Nicholas R Jennings, Ece Kamar, Isabel M Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C Parkes, Alex 'Sandy' Pentland, Margaret E Roberts, Azim Shariff, Joshua B Tenenbaum, and Michael Wellman. 2019. Machine behaviour. *Nature*, 568(7753):477–486.

Gayathri Rajendran, Bhadrachalam Chitturi, and Prabaharan Poornachandran. 2018. Stance-In-Depth Approach Deep Neural Approach to to Stance Stance Classification. *Procedia Computer Science*, 132(Iccids):1646–1653.

Martin Rajman and Martin Vesely. 2004. From Text to Knowledge: Document Processing and Visualization: a Text Mining Approach. pages 7–24. Springer, Berlin, Heidelberg.

Karthik Ramasubramanian and Abhishek Singh. 2017. *Machine Learning Using R*. Ml.

Andrew Ravenscroft. 2007. Promoting thinking and conceptual change with digital dialogue games. *Journal of Computer Assisted Learning*, 23(6):453–465.

Jan Recker. 2013. Scientific Research in Information Systems.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie Francine Moens. 2008. Language resources for studying argument. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 2613–2618.

Brian J Reiser. 2004. Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. Technical Report 3.

Antonio Reyes and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. In *Decision Support Systems*, volume 53, pages 754–760. North-Holland.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist.

Eric Ries. 2011. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.

Roman Rietsche and Matthias Söllner. 2019. Insights into Using IT-Based Peer Feedback to Practice the Students Providing Feedback Skill. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Henry L Roediger and Jeffrey D Karpicke. 2006. Test-Enhanced Learning Taking Memory Tests Improves Long-Term Retention. Technical report.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25*, pages 272–275.

Carolyn Rosé, Yi Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237–271.

Heinrich Roth. 1970. Pädagogische Psychologie des Lehrens und Lernens.

Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System System for Factual Knowledge. (Chi):1–13.

Victoria L. Rubin, Yimin Chen, and Lynne Marie Thorimbert. 2010. Artificially intelligent conversational agents in libraries. *Library Hi Tech*, 28(4):496–522.

Omar Sabaj, Paula Cabezas, Germán Varas, Carlos González-Vergara, and Álvaro Pina-Stranger. 2020. Empirical Literature on the Business Pitch: Classes, Critiques and Future Trends. *Journal of technology management & innovation*, 15(1):55–63.

D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144.

Steven L. Salzberg. 1997. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery*, 1(3):317–328.

Breno Santana Santos, Methanias Colaqo Junior, and Janisson Gois De Souza. 2018. An Experimental Evaluation of the NeuroMessenger: A Collaborative Tool to Improve the Empathy of Text Interactions. *Proceedings - IEEE Symposium on Computers and Communications*, 2018-June:573–579.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument Extraction from News. *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Jeremiah Scanlan. 2019. Auditing Predictive Policing. *Brigham Young University Prelaw Review*, 33.

Katharina Scheiter and Peter Gerjets. 2007. Learner control in hypermedia environments. *Educational Psychology Review*, 19(3):285–307.

Oliver Scheuer. 2015. *Towards adaptive argumentation learning systems*.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102.

Oliver Scheuer, Bruce M Mclaren, Frank Loll, and Niels Pinkwart. 2012a. Automated Analysis and Feedback Techniques to Support and Teach Argumentation: A Survey. *Educational Technologies for Teaching Argumentation Skills*, 2:71–124.

Oliver Scheuer, S Niebuhr, Toby Dragon, Bruce M McLaren, and Niels Pinkwart. 2012b. Adaptive Support for Graphical Argumentation – The LASAD Approach. *IEEE Learning Technology Newsletter*, 14(1):8–11.

Anuschka Schmitt, Thiemo Wambsganss, Andreas Janson, and Matthias Söllner. 2021. Towards a Trust Reliance Paradox ? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas.

Christoph Schneider, Markus Weinmann, and Jan Vom Brocke. 2018. Digital nudging: Guiding online user choices through interface design. *Communications of the ACM*, 61(7):67–73.

Sofia Marlena Schöbel, Andreas Janson, and Matthias Söllner. 2020. Capturing the complexity of gamification elements: a holistic approach for analysing existing and deriving novel gamification designs. *European Journal of Information Systems*, 29(6):641–668.

Julia E. Seaman, I. E. Allen, and Jeff Seaman. 2018. Higher Education Reports - Babson Survey Research Group. Technical report.

Dafna Shahaf and Eyal Amir. 2007. Towards a Theory of AI Completeness.

Priya Sharma and Michael Hannafin. 2007. Scaffolding in technology-enhanced learning environments. *Interactive Learning Environments*, 15(1):27–46.

Bayan Abu Shawar and Eric Steven Atwell. 2005. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, 10(4):489–516.

John Sinclair. 2005. Developing Linguistic Corpora: a Guide to Good Practice.

Paul Smolensky, Barbara Fox, Roger King, and Clayton Lewis. 1988. Computer-Aided Reasoned Discourse or, How to Argue with a Computer $|Taylor\&FrancisGroup$. In *R. Guindon (Ed.), Cognitive Engineering in the Design of Human-Computer Interaction and Expert Systems*, pages 109–162. Amsterdam.

Elliot Soloway, Mark Guzdial, and Kenneth E Hay. 1994. Learner-Centered Design The Challenge For WC1 In The Xst Century. *Interactions*, pages 36–48.

Donggil Song, Eun Young Oh, and Marilyn Rice. 2017. Interacting with a conversational agent system for educational purposes in online courses. *Proceedings - 2017 10th International Conference on Human System Interactions, HSI 2017*, pages 78–82.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78.

Christian Sonnenberg and Jan vom Brocke. 2012. Evaluation Patterns for Design Science Research Artefacts. pages 71–83. Springer, Berlin, Heidelberg.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from Old Man's View: Assessing Social Bias in Argumentation.

Christian Stab and Iryna Gurevych. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)(Oct. 2014), Association for Computational Linguistics, p.(to appear)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. *arXiv preprint*.

Karsten Stegmann, Christof Wecker, Armin Weinberger, and Frank Fischer. 2012. Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science*, 40(2):297–323.

Pontus Stenetorp, Sampo Pyysalo, and Goran Topi. 2012. BRAT : a Web-based Tool for NLP-Assisted Text Annotation. (Figure 1):102–107.

Ion Stoica, Dawn Song, Raluca Ada Popa, David Patterson, Michael W Mahoney, Randy Katz, Anthony D Joseph, Michael Jordan, Joseph M Hellerstein, Joseph E. Gonzalez, Ken Goldberg, Ali Ghodsi, David Culler, and Pieter Abbeel. 2017. A Berkeley View of Systems Challenges for AI.

Jong Hwan Suh. 2016. Comparing writing style feature-based classification methods for estimating user reputations in social media.

Patrick Suppes and Mona Morningstar. 1969. Computer-assisted instruction. *Science*, 166(3903):343–350.

Daniel D. Suthers. 2003. Representational Guidance for Collaborative Inquiry. In *Arguing to Learn*, pages 27–46.

Daniel D Suthers and Christopher D Hundhausen. 2001. European Perspectives on Computer-Supported Collaborative Learning. Technical report.

John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312.

Daniel Szopinski, T. Schoormann, and Dennis Kundisch. 2019. Because your taxonomy is worth it: Towards a framework for taxonomy evaluation. *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-basedmethods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, and Fakeeha Fatima. 2016. Text Mining : Techniques , Applications and Issues. 7(11):414–418.

Richard H. Thaler and Cass R. Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Heikki Topi. 2018. Using competencies for specifying outcome expectations for degree programs in computing: Lessons learned from other disciplines. *2018 SIGED International Conference on Information Systems Education and Research*.

Lisa Torrey and Jude Shavlik. 2012. Transfer Learning. *Encyclopedia of the Sciences of Learning*, pages 3337–3337.

Stephen E. Toulmin. 1984. *Introduction to Reasoning*.

Stephen E. Toulmin. 2003. *The uses of argument: Updated edition*.

Crystal Tse and Katie Roof. 2021. Coursera Prices IPO at Top of Range to Raise $519 Million - Bloomberg.

Tuure Tuunanen and Ken Peffers. 2018. Population targeted requirements acquisition. *European Journal of Information Systems*, 27(6):686–711.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Cathy Urquhart. 2017. Grounded Theory for Qualitative Research: A Practical Guide. *Grounded Theory for Qualitative Research: A Practical Guide*.

Muhammad Usman, Ricardo Britto, Jürgen Börstler, and Emilia Mendes. 2017. Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method.

Alper Kursat Uysal and Serkan Gunal. 2014. The Impact of Preprocessing on Text Classification. *Inf. Process. Manage.*, 50(1):104–112.

Suzanne Van Den Bosch. 2017. Automatic feature generation and selection in predictive analytics solutions. page 40.

Frans H Van Eemeren and Rob Grootendorst. 2016. *Argumentation, communication, and fallacies: A pragma-dialectical perspective*. Routledge.

C.J. van Rijsbergen. 1979. *Information retrieval*. Ph.D. thesis.

Maarten Vansteenkiste and Richard M Ryan. 2013. On Psychological Growth and Vulnerability: Basic Psychological Need Satisfaction and Need Frustration as a Unifying Principle.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *In Advances in neural information processing systems*, (Nips):5998–6008.

John Venable, Jan Pries-Heje, and Richard Baskerville. 2012. A Comprehensive Framework for Evaluation in Design Science Research. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7286 Lncs:423–438.

John Venable, Jan Pries-Heje, and Richard Baskerville. 2016. FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, 25(1):77–89.

Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, 39(2):273–315.

Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3):425–478.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. Annotating Argument Schemes. *Argumentation*, 35(1):101–139.

Jan vom Brocke, Wolfgang Maaß, Peter Buxmann, Alexander Maedche, Jan Marco Leimeister, and Günter Pecht. 2018. Future Work and Enterprise Systems. *Business and Information Systems Engineering*, 60(4):357–366.

Jan vom Brocke, Alexander Simons, Kai Riemer, Bjoern Niehaves, Ralf Plattfaut, and Anne Cleven. 2015. Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems*, 37(1):205–224.

Claudia von Aufschnaiter, Sibel Erduran, Jonathan Osborne, and Shirley Simon. 2008. Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching*, 45(1):101–131.

Ernst Von Glasersfeld. 1987. Learning as a constructive activity. *Problems of representation in the teaching and learning of mathematics*, 3(17):83–90.

Lev Semenovich Vygotsky. 1980. *Mind in society: The development of higher psychological processes*. Harvard university press.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A Review Corpus for Argumentation Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, CICLing 2014, pages 115–127, New York, NY, USA. Springer-Verlag New York, Inc.

D N Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. Argumentation Schemes for Presumptive Reasoning. L. Erlbaum Associates.

Douglas Walton and Fabrizio Macagno. 2015. A Classification System for Argumentation Schemes. *New Publisher: Ios Press*, 6(3):219–245.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.

Thiemo Wambsganss. 2021. Designing Adaptive Argumentation Learning Systems Based on Artificial Intelligence. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*.

Thiemo Wambsganss. 2022. Nudging Towards Self-Regulated Argumentation Learning: Designing Intelligent Argumentation Writing Support Systems. *Institute of Information Management at the University of St.Gallen Working Paper, St. Gallen, Switzerland*, pages 1–30.

Thiemo Wambsganss and Christian Engel. 2021. Using Deep Learning for Extracting User-Generated Knowledge from Web Communities. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*, pages 1–16.

Thiemo Wambsganss, Christian Engel, and Hansjörg Fromm. 2021a. Improving Explainability and Accuracy through Feature Engineering : A Taxonomy of Features in NLP-based Machine Learning. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas.

Thiemo Wambsganss, Sebastian Guggisberg, and Matthias Söllner. 2021b. ArgueBot: A Conversational Agent for Adaptive Argumentation Feedback. In *16th International Conference on Wirtschaftsinformatik*, Essen, Germany.

Thiemo Wambsganß, Leonie Haas, and Matthias Söllner. 2021. Towards the Design of a Student-Centered Question-Answering System in Educational Settings. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*.

Thiemo Wambsganss, Anne Höch, Naim Zierau, and Matthias Söllner. 2021a. Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design. In *Proceedings of the 16th International Conference on Wirtschaftsinformatik*, Essen, Germany.

Thiemo Wambsganss, Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. 2021b. AI-based Argumentation Tutoring – A Novel System Class to Improve Learners' Argumentation Skills. In *Annual Meeting of the Academy of Management (AOM). - Virtual Conference, 2021*, volume 2021, page 10233.

Thiemo Wambsganss, Tobias Kueng, Matthias Söllner, and Jan Marco Leimeister. 2021c. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA.

Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020a. Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach. In *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760, Dublin, Ireland. Association for Computational Linguistics.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020b. AL : An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA. Association for Computing Machinery.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020c. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*, Barcelona, Spain.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021d. Supporting Cognitive and Emotional Empathic Writing of Students. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4063–4077.

Thiemo Wambsganss and Roman Rietsche. 2019. Towards designing an adaptive argumentation learning tool. In *40th International Conference on Information Systems, ICIS 2019*, Munich, Germany.

Thiemo Wambsganss, Anuschka Schmitt, Thomas Mahnig, Anja Ott, Sigita Söllner, Ngoc Anh Ngo, Jerome Geyer-klingeberg, Janina Nakladal, and Jan Marco Leimeister. 2021e. The Potential of Technology-Mediated Learning Processes : A Taxonomy and Research Agenda for Educational Process Mining. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas.

Thiemo Wambsganss, Matthias Soellner, Kenneth R Koedinger, and Jan Marco Leimeister. 2022a. Adaptive empathy learning support in peer review scenarios. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Thiemo Wambsganss, Matthias Söllner, Andreas Janson, Ken Koedinger, and Jan Marco Leimeister. 2022b. Improving Argumentation Based on a Dynamic Feedback System Created with Machine Learning. *Under review at Information Systems Research*, pages 1–38.

Thiemo Wambsganss, Matthias Söllner, and Jan Marco Leimeister. 2020d. Design and Evaluation of an Adaptive Dialog-Based Tutoring System for Argumentation Skills. In *International Conference on Information Systems (ICIS)*, Hyderabad, India.

Thiemo Wambsganss, Florian Weber, and Matthias Söllner. 2021f. Design and Evaluation of an Adaptive Empathy Learning Tool. In *Hawaii International Conference on System Sciences*.

Thiemo Wambsganß, Florian Weber, and Matthias Söllner. 2021. Designing an Adaptive Empathy Learning Tool. In *16th International Conference on Wirtschaftsinformatik*, Essen, Germany.

Thiemo Wambsganss, Florian Weber, Tobias Staufenberg, Leopold Bott, and Matthias Söllner. 2020e. Systematische Entwicklung eines Lerntools zur Erhöhung der Argumentationsfähigkeiten von Studierenden. In *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany.

Thiemo Wambsganss, Rainer Winkler, Pascale Schmid, and Matthias Söllner. 2020f. Designing a Conversational Agent as a Formative Course Evaluation Tool. In *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany.

Thiemo Wambsganss, Rainer Winkler, Pascale Schmid, and Matthias Söllner. 2020g. Unleashing the Potential of Conversational Agents for Course Evaluations: Empirical Insights from a Comparison with Web Surveys. In *Twenty-Eighth European Conference on Information Systems (ECIS2020)*, pages 1–18, Marrakesh, Morocco.

Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020h. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.

Wenting Wang, Deeksha Arya, Nicole Novielli, Jinghui Cheng, and Jin L.C. Guo. 2020. ArguLens: Anatomy of Community Opinions on Usability Issues Using Argumentation Models. *Conference on Human Factors in Computing Systems - Proceedings*.

Joe H. Ward. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236.

Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Lernaing : A Taxonomy of Conversational Agents in Education. In *Forty-Second International Conference on Information Systems*, pages 1–17, Austin, Texas.

Jane Webster and Richard T Watson. 2002. Analyzing the past to prepare for the future : Writing a literature review Reproduced with permission of the copyright owner . Further reproduction prohibited without permission . *MIS Quarterly*, 26(2):xiii–xxiii.

World Economic Forum WEF. 2018. The Future of Jobs Report 2018. Technical report.

Robert J. Weijers, Björn B. de Koning, and Fred Paas. 2020. Nudging in education: from theory towards guidelines for successful implementation. *European Journal of Psychology of Education*, pages 1–20.

Armin Weinberger and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers and Education*, 46(1):71–95.

Armin Weinberger, Karsten Stegmann, and Frank Fischer. 2010. Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human Behavior*, 26(4):506–515.

Markus Weinmann, Christoph Schneider, and Jan vom Brocke. 2016. Digital Nudging. *Business and Information Systems Engineering*, 58(6):433–436.

Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang. 2010. *Fundamentals of Predictive Text Mining*.

Radboud Winkels, Jochem Douw, and Sara Veldhoen. 2013. Experiments in automated support for argument reconstruction. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law - ICAIL '13*, page 232.

Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Conference on Human Factors in Computing Systems - Proceedings*.

Rainer Winkler and Matthias Söllner. 2018. Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. *Academy of Management Proceedings*.

Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2021. Enhancing problem-solving skills with smart personal assistant technology. *Computers and Education*, 165.

David Wood, Jerome S. Bruner, and Gail Ross. 1976. The Role Of Tutoring In Problem Solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

Laura Wood. 2016. Artificial Intelligence (AI) Market By Technology (Machine Learning, Natural Language Processing (NLP), Image Processing, And Speech Recognition), Application & Geography - Global Forecast To 2020. Technical report.

Anita Woolfolk Hoy. 2000. Educational psychology in teacher education. *Educational Psychologist*, 35(4):257–270.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. *Conference on Human Factors in Computing Systems - Proceedings*, 2017-May:3506–3510.

Junyi Xu, Li Yao, and Le Li. 2015. Argumentation based joint learning: A novel ensemble learning approach. *PLoS ONE*, 10(5):1–21.

Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. 2021. From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI. pages 1–73.

Yiming Yang and Jan O Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, Icml '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1:180–189.

Robert K Yin. 2015. *Qualitative research from start to finish*. Guilford publications.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing.

Hannes Zacher and Michael Frese. 2018. Action regulation theory: Foundations, current knowledge and future directions. In *The SAGE handbook of industrial, work & organizational psychology: Organizational psychology, Vol. 2, 2nd ed.*, pages 122–144. Sage Reference.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016a. ArgRewrite: A Web-based Revision Assistant for Argumentative Writings. 2016:37–41.

G Zhang, S Purao, Y Zhou, and H Xu. 2016b. Argument detection in online discussion: A theory based approach. *AMCIS Proceedings*, (Yates 1996):1–10.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393.

Naim Zierau, Thiemo Wambsganss, Andreas Janson, Sofia Schöbel, and Jan Marco Leimeister. 2020. The Anatomy of User Experience with Conversational Agents: A Taxonomy and Propositions of Service Clues. In: International Conference on Information Systems (ICIS). Anatomy of User Experience with Conversational Agents The Anatomy of User Experience with. In *International Conference on Information Systems (ICIS)*, Hyderabad, India.

Barry J. Zimmerman. 1989. A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3):329–339.

Barry J Zimmerman and Dale H Schunk. 2001. *Self-regulated learning and academic achievement: Theoretical perspectives*. Routledge.

Verena Zimmermann and Karen Renaud. 2021. The nudge puzzle: Matching nudge interventions to cybersecurity decisions. *ACM Transactions on Computer-Human Interaction*, 28(1).

# A Appendix

## A.1 Guidelines for the Annotation of Argumentation Structures in Student-Written Peer Reviews

### A.1.1 Introduction

Argumentation mining is an interdisciplinary research area that includes philosophy, psycholinguistics, and computer science for building argumentation models and automated methods for identifying arguments in written texts. These tools not only provide new opportunities educational applications such as intelligent writing support, information retrieval platforms, or automated assessment tools, but also open new possibilities for improving current legal information retrieval platforms or policy modeling platforms (Stab and Gurevych, 2017).

In this work, a corpus of one thousand annotated student reviews was created and then the argumentation was evaluated in terms of how helpful the argumentation was, how high quality it was, and how good the argumentation structure was. How helpful an argument was was rated on how understandably arguments were expressed and how meaningful they seemed to the reader. Quality was rated according to how strong an argument was, i.e., if arguments were supported by one or more meaningful premises. The more meaningful premises, the stronger an argument. Structure was evaluated by how easily and logically relationships could be established between the components of an argument. Students each argumentatively evaluated a written paper by another student. In doing so, they addressed the strengths and weaknesses of a paper and then suggested improvements for revision.

The ability to formulate persuasive arguments is a crucial aspect of acquiring writing skills. On the one hand, well-defined arguments are the basis for convincing the reader, on the other hand, good argumentation skills are essential for analyzing different positions in general decision making (Stab and Gurevych, 2014a). In the case of student reviews, sound argumentation gives a good impression of how well the underlying student work has been handled. It uses expressive arguments to show the strengths and weaknesses to the student who wrote the paper and provides argumentatively supportive suggestions for improvement. Rules were established for the annotation of the student reviews, which build on the basis of the rules created by (Stab and Gurevych, 2014a). The rules were adapted because these are German-language reviews and not English-language persuasive essays.

### A.1.2 Argument components

An argument consists of several statements. In its simplest form, it contains a statement (claim) supported by a single premise (premise) (Toulmin, 2003; Walton et al., 2008). The statement represents a controversial proposition that the author is trying to convince the reader of. It is usually a proposition or assumption and should not be accepted by the reader without additional support (Stab and Gurevych, 2014a). Premises can thereby not only support an opinion, but also attack it, which is used as a stylistic device or can also illustrate uncertainty in the argument. There can also be more complicated constellations of statements and premises, in which a statement is supported by several different premises or by a chain of premises, in which each premise is in turn supported by another premise. Similarly, an opinion may be supported by one premise and attacked by another, or supported by a premise which is in turn attacked by another premise. The simplest form can look like this:

*opinion* because *premise*.

### A.1.3 Limits of argumentation components

Argument components do not necessarily cover a complete sentence, because often a sentence can also consist of several argument components, which are annotated separately. For example, a sentence can consist of a statement and one or more premises. Likewise, a sentence can begin with so-called phrases such as "In my opinion...". These are not relevant for the content of the argumentation component as long as the rest still makes a complete sentence, but more about this in the following established rules.

#### A.1.3.1 Completeness rule

An argument component should always include a point of view, which can also be a complete sentence. In English, you can check whether an annotated component represents a complete point of view by prefixing the statement with the phrase "It is true that, *claim*" and the sentence still remains grammatically correct. If this is fulfilled, then the annotation is correct according to the completeness rule (Stab and Gurevych, 2014a) In German, and especially in the case of reviews, the rule cannot be applied one-to-one, since the German sentence structure differs from the English sentence structure and the wording in reviews is also different. Thus, as soon as you put the sentence part "It is true that *opinion*" in front of the opinion, the sentence structure of the opinion must be

changed so that it still remains grammatically correct. The following examples illustrate this. The sentence part in the square brackets contains the correct boundaries of the argument component of the annotation.

English example: *"[The customer Process is not described in sufficient detail]."*

German translation: *„[Der Kundenprozess ist nicht genügend detailliert beschrieben].“*

Rule application in English: It is true that, "*[ the customer Process is not described in sufficient detail]."*

Rule application in German: *Es ist wahr, dass „[der Kundenprozess ist nicht genügend detailliert beschrieben].“* English example: *"[ You'd have to justify that a bit more] because [that's a very important point that has a big impact on your cost structure]."*

Rule application in English: *It is true that, "[ you'd have to justify that a bit more]," / It is true that, [that's a very important point that has a big impact on your cost structure].*

German example: *„[Das müsstest du noch etwas genauer begründen], da [das ein sehr wichtiger Punkt ist, der einen hohen Einfluss auf deine Kostenstruktur hat].“*

Rule application in German: *Es ist wahr, dass „[das müsstest du noch etwas genauer begründen],“ / Es ist wahr, dass [das ein sehr wichtiger Punkt ist, der einen hohen Einfluss auf deine Kostenstruktur hat].“*

The first example illustrates that the German sentence must be rearranged to keep the sentence grammatically correct. The sentence should read as follows: It is true that "[the customer process is not described in sufficient detail]." The second example shows that there only the first part of the sentence must be grammatically adjusted. So that the rule applies. The sentence would have to read as follows: It is true that "[you would need to justify this a little more],".

German rule: If "It is true that" can be placed in front of the argument component and the sentence structure of the argument component can be rearranged with the same words in such a way that the sentence remains grammatically correct, then the annotation is fulfilled according to the completeness rule.

#### A.1.3.2 Relevance rule

All words that are relevant for the argumentation component must be included in it. This means that all subordinate clauses relevant to the content must be included in

the annotated component. Temporal information such as "in the past", "recently" or "nowadays", which can be at the beginning of a sentence, must also be included in the annotated component, otherwise the content cannot be understood (Stab and Gurevych, 2014a).

English example: *"[ At the moment, I still see a weakness in your business model]."*

German translation: *„[Momentan sehe ich noch eine Schwäche in deinem Business Modell]."* This rule can be adopted into German, since here it is further facilitated by the fact that such information is in most cases not separated from the rest of the sentence by any punctuation mark. Furthermore, what is meant here is that the business model represents a weakness before a revision and not that it represents another weakness, as one would otherwise interpret the word "noch".

#### A.1.3.3 Shell language rule

In English, shell language has no relevance to the context of an argument component because the component is nevertheless grammatically correct and thus irrelevant to it. Examples of such shell words include "For example," "According to the previous fact," "As can be seen," "Another important point which contributes to my argument is that", "In this context", "because", "though", "in my view", "i think", "also", "furthermore", etc. In the exceptional case that shell words are relevant to the context of an argument component, they should be included in the component. Examples are: "I do not agree that", "I disagree with the view that", "i don't think", "i do not believe", etc. (Stab and Gurevych, 2014a). In German recessions, this rule cannot always be adopted one-to-one either, since after such shell words the argumentation component is not always grammatically correct, unless you add another word like "es".

Translating the English shell words into German gives the following examples of shell words: "For example", "According to the previous fact", "As you can see", "Another important point which supports my argument is that", "In the context", "because", "since", "however", "nevertheless", "In my opinion", "I think", "also", "Furthermore", etc. The translation of the shell words relevant to the context of an argument component are: "I disagree with that", "I disagree with the view that", "I don't think that", "I don't believe that". In the following, I show an example where the shell word can be omitted and the German sentence remains grammatically correct and then an example where the sentence remains grammatically incorrect by omitting the shell word without adding another word.

English example: *"I think [the description of the short-characteristics of your company is a bit too long]. "*

Deutsche Übersetzung: *„Ich denke [die Beschreibung zu den Kurzcharakteristika deines Unternehmens ist etwas zu lang geraten]. “*

English example: *"Furthermore, [ it is clear that the business model was developed from a consumer perspective]."*

Deutsche Übersetzung: *„Des Weiteren [ist klar ersichtlich, dass das Geschäftsmodell aus Konsumentensicht entwickelt wurde]. “* As can be seen in the second example, the German argumentation component as a stand-alone sentence so grammatically does not make sense without adding the word "es". The argument component is grammatically correct as follows: "Furthermore, [it is clearly evident that the business model was developed from the consumer's point of view]." You can rephrase the sentence as follows according to the completeness rule so that the grammar remains correct: "Furthermore [it is clearly evident that the business model was developed from a consumer perspective]." I.e. here shell word "furthermore" should be included in the argumentation component and the sentence should be annotated as follows: "Furthermore [it is clearly evident that the business model was developed from the consumer's point of view]." An example of shell words that are Relevant to the context of an argument component is as follows:

English example: *"I do not believe that [good IT skills are a USP]."* Deutsche Übersetzung: *„Ich glaube nicht, dass [gute IT-Kenntnisse ein USP sind]. “*

In this case, one can clearly see that without the phrases, the context of the argument component says exactly the opposite. Therefore, the phrases are included in the annotation here and the sentence is thus annotated as follows: "[I don't believe that good IT skills are a USP]." For these reasons, the following German rule can be established:

German rule: If an argumentation component can be rearranged according to the completeness rule and the context of the component is not changed by this, then shell words or phrases should not be included in the annotation. If the context of the argument component is changed by leaving out shell words, then they should be included in the annotation. Additionally, shell words in English as well as in German serve as a good indicator for identifying argument components.

#### A.1.3.4 Separation rule

A sentence may only be fully annotated if the corresponding sentence does not contain a

conclusion and shell words between different opinions. I.e. a sentence may not be split into two argument components if there are several complete opinions. Only if one opinion is the reason for another opinion, the sentence may be split into two or more argument components. It is important not to separate sentences with multiple opinions connected by words such as "and", "or", and punctuation marks as in a bulleted list, and which do not contain a conclusion between different opinions. This can often occur when one sentence contains multiple reasons for an opinion from another sentence. In this case, the different reasons (premises) from one sentence must be annotated as one argument component. This also applies to conditional sentences, which contain a condition under which something occurs (Stab and Gurevych, 2014a). An indicator of such sentences are words like "if", or "in that". The subordinate clause formulates the condition that must be met in order for the consequence stated in the main clause to be realized.

English example: *"This is due to the fact that, [you could add more examples and you should describe them in more detail]."* German translation: *„Dies ist dem Grund geschuldet, dass [du noch viele Beispiele ergänzen könntest und du diese noch genauer beschreiben solltest]."*

In this example, the argument component contains two premises: "You could add more examples" and "You should describe these in more detail". These are connected by the word "and". I.e. this rule can be adopted for German reviews, which means that these two premises should be combined into one argumentation component 3.5 Punctuation rule

Punctuation marks that come at the end of an argument component must not be included in the annotation (Stab and Gurevych, 2014a).

English example: *"[EasyHealth is a business model that largely meets the vision of WellBeing]."* Deutsche Übersetzung: *„[Bei EasyHealth handelt es sich um ein Geschäftsmodell, das größtenteils die Visionen des WellBeings erfüllt]."*

This rule can also be transferred one-to-one for German reviews, since English and German grammar do not differ there.

#### A.1.3.5 Interpretation rule

Sentences which only represent an unambiguous statement through an interpretation should be annotated as an argumentation component. In particular, statements that explain what the author of the work has done and can be interpreted as a weakness;

statements in which the author writes that he has not understood something and can be interpreted to the effect that these should be explained in more detail; or questions that suggest that an explanation or an important point has been forgotten by the author of the work. It must be remembered that these statements or questions always fall under the points of strength or weakness of the work. I.e. it is assumed with the statement or question that a point from the work of the author represents a weakness or that an important point is simply missing in the eyes of the evaluator or was inaccurately explained. These statements or questions may be supported by premises or, in some cases, may be written alone without a premise. All of these sentences suggest an opinion, if interpreted correctly. Such sentences occur frequently in a review, making this rule particularly relevant to this corpus. Indicators of such sentences are expressions such as "You have," "I don't understand," "I'm not quite clear," or a punctuation mark such as "?". In these cases, as already mentioned, an interpretation is often necessary to present a statement. If an unambiguous statement can be interpreted, this argument component is included in the annotation. Premises can strengthen the unambiguity of a statement.

Deutsches Beispiel: *„[Du hast gute IT-Kenntnisse als USP aufgeführt]. Ich glaube nicht, dass [gute IT-Kenntnisse ein USP sind].“*

Deutsches Beispiel: *„[Ich verstehe nicht, was du mit Steuerung und Modifizierung der App meinst].“*

Deutsches Beispiel: *„[Was ist mit den Kosten für deine Entwickler?]“.*

These examples show that the previously described statements or questions can each be interpreted as an opinion, provided that one considers that these statements and question were each listed under the item weakness. Therefore, such sentences should each be annotated as an argument component and, as in these examples, as an opinion. Only the first example is additionally supported by a premise. In the same way, however, such examples can also represent premises, especially if suggestions for improvement have been written by the evaluator.

German example: *„[Du hast gute IT-Kenntnisse als USP aufgeführt]. Ich glaube nicht, dass [gute IT-Kenntnisse ein USP sind]. [Du solltest die IT-Kenntnisse in der SWOT-Analyse unter Stärken aufführen].“*

German example: *„[Ich verstehe nicht, was du mit Steuerung und Modifizierung der App meinst]. [Ich empfehle dir genauer zu beschreiben, was du mit Steuerung und Modifizierung der App meinst].“*

German example: *„[Was ist mit den Kosten für deine Entwickler?]. [Du solltest die*

*Kosten für Entwickler auf jeden Fall mit in deine Kostenaufstellung mit reinnehmen].“*

These examples illustrate that the argument components previously annotated as opinions are now supported by the subsequent improvement proposals and should be annotated as premises.

#### A.1.3.6 Grammar rule

Sentences that are constructed like bullet points, in which grammatically important words are omitted but still represent a clear statement or premise, should be treated as if those words were present. This occurs frequently in reviews when examples are given. Likewise, sentences that are grammatically incorrect should be annotated as argumentative. Statements or premises that are split into several sentences and only together make a meaningful argument component are annotated as one argument component.

German example: *„[Klar verständliches Geschäftsmodell].“*

German example: *„[Es gibt extra ein BMC für Two-Sided Markets, das halte ich aber für unübersichtlich unnötig].“*

In the first example, the subject and predicate are missing. Correctly, this sentence should read as follows: This is a clearly understandable business model. The context is clearly understandable despite the missing subject and predicate. Therefore, this statement should be annotated as an argument component. In the second example, the context becomes clearer if you formulate the sentence grammatically correctly as follows: [I consider the extra BMC for Two-Sided Markets to be confusing and unnecessary]. The statement is thus unambiguous and the argument component is annotated.

#### A.1.3.7 Chapter name rule

In reviews, opinions may be divided by chapter and misinterpreted without the chapter name. Chapter names can be relevant to the context of an argument component and should be annotated with in such cases.

German example: *„[Revenue Streams Die sind gut beschrieben].“*

German example: *Value Proposition Ich finde [die Value Proposition ist umfassend und ausführlich beschrieben].*

The first example shows that the chapter name is relevant to the context of the statement and should therefore be annotated. In the second example, on the other hand, the chapter name is not relevant to the context of the statement and should therefore not be included in the annotation.

### A.1.4 Annotation of opinions and premises in reviews

A statement in the context of this work is directly related to the overall criteria of how the review should be written by the students. These criteria are that in each case the strengths and weaknesses should be highlighted to the work to be evaluated, as well as suggestions for improvement should be attached. That is, a statement always suggests a strength, weakness, or suggestion for improvement. Thus, one should always be able to formulate the following: A strength is that [opinion], A weakness is that [opinion] and a suggestion for improvement is that [opinion].

German example: *Eine Stärke ist, dass „[du dein Geschäftsmodell sehr detailliert erläutert hast].“*

German example: *Eine Schwäche ist, dass „[der Kundenprozess nicht genügend detailliert beschrieben ist].“*

German example: *Ein Verbesserungsvorschlag ist, dass [du die IT-Kenntnisse in der SWOT-Analyse unter Stärken aufführen solltest].“*

Premises, on the other hand, are a reason why an argument component (an opinion or other premise) is supported or attacked. That is, a premise represents a justification or a refutation of an argument component in order to convince the reader of the truth or falsity of an opinion or other premise.

German example: *„[Es ist sehr schwierig eine solche Lösung zu vermarkten], da [sie viel zu viele verschiedene Features hat]“*

### A.1.5 Argumentative relations

Argument components form a tree structure of an argument through relationships (Stab and Gurevych, 2014a). An argument always contains a statement, which can be supported or attacked by one or more premises. In turn, individual premises can also be supported or attacked by a chain of other premises. ”Supported” can be replaced by ”attacked” at any point. Likewise, the argumentation tree can be extended infinitely by other premises

and relations.

## A.2 Guideline for the Annotation of Argument Structure and Quality in Business Pitches

### A.2.1 Introduction

Argumentation is a ubiquitous basis of our daily communication and thinking. In general, it is intended to increase or decrease acceptance of a controversial point of view (Eemeren et al., 1996). Logical, structured arguments are a necessary prerequisite for persuasive conversations, general decisions, and drawing accepted conclusions (Kuhn 1992). Based on Aristotle's studies, the ability to formulate persuasive arguments is recognized as the basis for convincing an audience of new ideas; it plays an important role in strategic decision making and the analysis of different points of view.

Argumentation mining is an interdisciplinary research area from computational linguistics that involves the creation of argumentation models and automated methods for identifying arguments in written texts. These models not only provide novel opportunities for educational applications such as intelligent writing support, but also open up new possibilities for improving student argumentation through adaptive learning tools. However, an important prerequisite for the development of Natural Language Processing (NLP) methods that can identify argument components and argumentative relations in written texts is the availability of annotated corpora. However, current annotated argumentation corpora are usually limited to certain domains, e.g., legal documents (Mochales and Moens, 2008), newspapers and court cases (Chris Reed Raquel Mochales Palau and Moens, 2008), or online debates (Cabrio and Villata, 2012). In Educational Texts, there are currently two annotated corpora of persuasive student essays by Stab and Gurevych (2014a) and Stab and Gurevych (2017). However, these corpora are annotated in English and are not derived from a specific teaching-learning scenario.

Because of this requirement and the complex structure of argumentative discourse, "automatic recognition of arguments has been little studied until 2008" (Chris Reed Raquel Mochales Palau and Moens, 2008). The goal of this work is to formulate a suitable guideline to help annotate the structure and quality of argumentation and its components in persuasive business pitches.

#### A.2.1.1 Arguments in a Nutshell

An argument consists of several statements. In its simplest form, it contains an statemment/claim supported by at least one premise (Toulmin, 2003). The claim represents a controversial statement with which the author attempts to persuade the reader. This usually represents a suggestion or assumption and should not be accepted by the reader without additional support. The characteristic feature distinguishes arguments from statements in which the conclusion is a true non-controversial statement (e.g., an event that occurred in the past). The second component of an argument, the premise supports the plausibility of the claim. This is usually added by the proponent (writer) to convince the reader of the claim.

Considering the simplest form of an argument, a premise can be seen as a justification for the claim, while more complex argument structures can also contain premises that aim to refute a claim. These more complex structures are basically graphs that connect premises and claims through different relationships.

Even such a simple form of an argument can be expressed in many different ways in written text.

In the further course of this guideline, a sound basis for the annotation of arguments in business pitches will be created using definitions, rules and concrete examples. The next chapter takes a closer look at business pitches as a form of argumentation and discusses their special features and structure. This is followed by the annotation process.

#### A.2.1.2 Persuasive Business Pitches

This annotation guideline is built based on a dataset of persuasive business pitches. These pitches usually have a certain structure, which is briefly explained in this section. Typically, a pitch begins with an introduction that includes a brief description of the business idea. This basic business idea is referred to below as the Major Claim.

The actual arguments that either support (support) or attack (attack/challenge) the major claim are presented after the introduction or after the major claim. At the end of the Business Pitch, there can be a summary of the supporting arguments (claims/premises), a reiteration of the business idea (major claim), or a concluding statement. The annotation structure follows a tree structure. The root node of the tree is the Major Claim, which mostly occurs in the Introduction. The Major Claim is supported by Claims, and the individual Claims are in turn supported by Premises.

Premises can either support (support) or attack (attack/challenge) claims. In Example 4,

both Claim 1 and Claim 2 are supported by premises. The premise supporting claim 1 is additionally supported by another premise. After describing the argumentation structure, the next section gives a brief overview of the annotation process and then describes each annotation step individually in the following sections.

#### A.2.1.3 Annotation process

The previous sections described in detail the argumentation structures that will be annotated in this study. For the annotation of these structures, the six-step process will be clearly presented in the following.

**1. Reading the entire Business Pitch and then annotating the Major Claims:**

In the first step, the Business Pitch is read in its entirety so that an initial understanding of the business model is given. Then the major claim, i.e. the business idea, should be annotated. This is usually found at the beginning or end of the text. Concrete examples of this are explained later in this guideline.

**2. Annotation of the claims:**

In the second step, we reread the text, marking each claim as the central component of an argument that either supports or attacks the Major Claim. A claim can thus be seen as a reason or assertion for the Major Claim.

**3. Annotation of the premises:**

In the third step, the premises for the claims are annotated. These are the reasons given by the author for supporting or attacking the claim. A premise can be a statistic, intuition, invented instance, analogy, or statement.

**4. Annotation of argumentative relations of the argumentation components**

The fourth step involves the annotation of the argumentative relations of claims and premises. Here we annotate the support and attack (support or challenge) relations that exist between argument components to identify the structure of the arguments. **5. Annotate the quality of the argumentation components**

In the fifth step, the quality of the argumentation components is evaluated by ranking major claims, claims and premises on the basis of various dimensions. The most important dimensions are specificity and evidence. Specificity refers to the limited scope of a claim. Evidence describes how well the supporting components describe the component to be

supported. After all argumentation components have been evaluated according to the present scheme, the persuasiveness can be assessed. The persuasiveness is derived from the scores of the individual dimensions. If a major claim, a claim or a premise has low scores in specificity and evidence (strength of evidence), the persuasiveness is to be rated as low. If individual dimensions have higher scores and individual dimensions have lower scores, the persuasiveness is to be rated as neutral or medium. If all dimensions have a high score, the argumentation component has a high persuasiveness (score).

**6. Annotation of the overall quality of the business pitch**

After the quality of the individual argumentation components has been evaluated, the overall quality of a pitch will finally be assessed. The strength, clarity and generation of interest are evaluated in turn (cf. Tables 9, 10, 11). The procedure is similar to that for the argumentation components, except that in this case the argumentation is evaluated as a whole.

### A.2.2 Argument components

An argument consists of several statements. In its simplest form, it contains a statement (claim) supported by a single premise (premise) (Toulmin, 2003). The statement represents a controversial proposition that the author is trying to convince the reader of. It is usually a proposition or assumption and should not be accepted by the reader without additional support(Stab and Gurevych, 2014a). In this context, premises can not only support an opinion, but also attack it, which is used as a stylistic device or can also illustrate uncertainty in the argument. The simplest form could be as follows:

The argument components of an annotation in business pitches consist of three argument components. First is the Major Claim (MC), which is the main statement of the pitch that reflects the fundamental idea of the business model (Carlile et al., 2018). The following statement shows an example of a Major Claim:

*"The new platform WeMatch connects users via existing dating apps (partners/ co-operations) with potential locations for a first meeting."*

The major claim thus briefly presents the basic idea in one sentence. The argument can follow one of three strategies, ethos (appeals to the author's credibility, pathos (appeals to the reader's emotions), and logos (author uses facts and evidence). The idea can be supported by one of the three strategies. The three strategies can be applied to both Major Claims and Claims. 1st example Major Claim (MC): Basic idea (main message).

Idea 1: *The presented platform - Lokker - will serve to link the demand for storage space with owners of vacant commercial space.*

Idea 2: *Wouldn't it be great if you could order photos directly from your cell phone and surprise your beloved wife or good friends with great pictures for Christmas? FOS makes it possible.*

Idea 3: *The "Grillsational App", based on an interactive platform where you can interact with the local grilling comm*unity - not only individuals but also businesses.

The second argument component is a claim, in the form of a controversial statement that readers should not accept without additional support (Carlile et al., 2018). To be able to convince the reader of the statements regarding the idea, premises are used. Premises are moving reasons to convince the reader of the idea. Premises can take two different forms, support or challenge. The support supports a claim or major claim.

*"Around a thousand square meters of commercial space has been vacant since the new building was built in 2016, which could bring in rental income of 200,000CHF annually."* This example would support the claim or major claim (idea 1). In addition to support premises, challenges can also be used in a pitch to draw attention to challenges relating to the idea. This is not considered a disadvantage, but a positive criterion. To make premises easier to spot there are some signal words that can be looked at. These include words such as "therefore", "therefore", "by", "thereby", "since", "because", "in order to", "so", "so that", "whereby". The words are mostly adverbs and can indicate premises, but are not a guarantee of premises and must be carefully examined in certain situations. 2. example claim (C): controversial statements that should not be accepted by readers without additional support.

3. example premise (P): reasons to convince the listener (reader) of the idea consists of two components challenge and support.

3.1 Support: one reasoning component supporting another one

3.2 Challenge: an argumentation component that shows challenges

Premises can be of different types, for example, statistics, explanation, description, invented instance, or analogies (Carlile et al., 2018). For example, a statistic is,

*"In this regard, MyCareerCoach benefits from strong recurring demand, as more than 400,000 students in Germany are considered for higher education each year. "* Examples of premise types:

**instance**: This means that environmentally harmful microplastics no longer enter the waterways. And plastic waste from packaging can be reduced to a minimum.

**Analogy**: Product tests such as those by K-Tipp are only a snapshot, while blog posts in turn only reflect the opinions of individual people.

**Explanation**: This prevents the network from placing a lot of advertising. At the same time, the payment barrier serves as a filter and quality control for members.

**Description**: Services can all be booked and paid for simultaneously and centrally.

**User-centric example:** when planning a special movie night, the customer can purchase additional services and products to go with the selected movie.

### A.2.3 Relationships of argumentation components

Argument components form the tree structure of an argument through relationships (Stab and Gurevych, 2017). An argument always contains a statement, which can be supported or attacked by one or more premises. In turn, individual premises can also be supported or attacked (challenged) by a chain of other premises. Each argument can thus consist of 3-4 levels. The first level is the Major Claim, which contains the main statement. In the second level there is a claim (assertion), which supports the first level. Premises form the second level, premises can include for example statistics, definitions or real examples. At the fourth level, there may be additional premises that provide additional support for level 3 (Carlile et al., 2018).

Having explained the tree structure in principle, we link the argument components with argumentative relations to build the tree structure of each argument. An argumentative relationship is a directed link between two argument components with a specific source and target component. Such a relationship indicates either that the source component is a support (support relationship) or an attack/challenge (attack relationship/challenge) for the target component (Stab and Gurevych, 2017). However, in this Guideline we assume that there will not be a classic Attack, since the students will not refute their own ideas in a short business pitch. Therefore, in this Guideline, we view the attack as a challenge, i.e., an indication of future challenges that might be related to the project or idea. Basically, each initial component of an argumentative relation represents a premise. By annotating argumentative relationships, we identify for each premise the goal to which it belongs and recognize whether the premise supports (Support) or attacks (Attack/Challenge) the goal. Note that the target can always be a claim or another premise. For the claims and the premises, the argument structure is built as follows (Stab and Gurevych, 2017).

Steps: 1. choose a claim 2. link each premise to the claim it supports (Support) or attacks (Attack / Challenge).

Following this process ensures that the argument structure is a tree and that each premise is associated with exactly one argument component (either Claim or Premise). Note also that in some cases the support (or attack/challenge) of a single premise may be weak and not obvious. However, when it is combined with another premise, the persuasive power can become stronger. This happens, for example, when the author uses a specific event or example to justify or refute a point of view. Therefore, it may sometimes be necessary to consider premises in combination to identify the proper target of each premise. In the next sections, the annotation of support and attack relationships is explained in detail using several examples(Stab and Gurevych, 2017).

#### A.2.3.1 Annotation of support relationships (Support)

A support relationship (support) between two argument components indicates that the source component provides a reason or justification for the target relationship. In the following examples, the claim is underlined, the premise is underlined in waves, and the argument components are enumerated by quotation marks. Consider the following simple example:

Simple example: "[Until now, it has been difficult for Swiss companies to rent out their storage space]1 because [this involves a great deal of effort (legal, administrative, time)]".

In this example, there are only two argument components, the first of which occurs as a claim and the second as a premise. The indicator "because" signals that the second component is a justification for the first component. Therefore, the second argumentation component is linked to the first one via a support relation (Support).

### A.2.4 Rules for annotation of argumentation components

Argument components do not necessarily cover a complete sentence, because often a sentence can also consist of several argument components, which are annotated separately. For example, a sentence may consist of a statement and one or more premises. Likewise, a sentence can begin with so-called phrases such as "In my opinion...". These are not relevant to the content of the argument component, as long as the rest still makes a complete sentence. The following rules should clarify this.

#### A.2.4.1 Completeness rule

An argument component should always include a point of view, which can also be a complete sentence. In English, one can check whether an annotated component represents a complete position by placing the clause "It is true that, *claim*" in front of the statement and the sentence still remains grammatically correct. If this is satisfied, then the annotation is correct according to the completeness rule (Stab and Gurevych, 2017).

For the example of business pitches, this rule cannot be applied one-to-one, since the German sentence structure is different and the formulations are also different. Thus, as soon as the sentence part "It is true that *opinion*" is placed in front of the statement, the sentence structure of the statement must be changed so that it still remains grammatically correct.

Rule: *If "It is true that" can be placed before the argument component and the sentence structure of the argument component can be rearranged with the same words in such a way that the sentence remains grammatically correct, then the annotation is fulfilled according to the completeness rule.*

Example: Buying and changing tires is an unpleasant activity. It consumes time and money and does little in terms of convenience. It is true that buying and changing tires is an unpleasant activity. It is true that it consumes time and money and can do little in terms of convenience.

#### A.2.4.2 Relevance rule

Rule: All relevant words must be included in the argument component. This means that all subordinate clauses relevant to the content must be included in the annotated component. Temporal information such as "in the past", "recently" or "nowadays", which can be at the beginning of a sentence, must also be included in the annotated component, otherwise the content cannot be understood (Stab and Gurevych, 2017).

Example: Nowadays, there are more singles with a disproportionately high income and women or mothers who have a part-time job. Accordingly, the old role model of man and woman is considered outdated.

#### A.2.4.3 Shell language rule

In English, shell language has no relevance to the context of an argument component, since the component is nevertheless grammatically correct and thus irrelevant to it. Examples of shell words include : "For example,", "According to the previous fact,", "As can be seen", "Another important point which contributes to my argument is that", "In this context", "because", "though", "in my view", "i think", "also", "furthermore",.... In the exceptional case that shell words are relevant to the context of an argument component, they should be included in the component. Examples are: "I do not agree that", "I disagree with the view that", "i don't think", "i do not believe", ... (Stab and Gurevych, 2017).

In business pitches, this rule also cannot be adopted completely in every case. After shell words, the argument component is not always grammatically correct unless another word like "it" is added. Translating the English shell words into German results in the following examples of shell words: "For example", "According to the previous fact", "As you can see", "Another important point which supports my argument is that", "In the context", "because", "since", "however", "nevertheless", "In my opinion", "I think", "also", "Furthermore", etc.

Example: Furthermore, [cooking diaries can be shared between friends via the WMF app].

The argument component makes no grammatical sense as a stand-alone sentence like this without adding the word "it". The argument component is grammatically correct as follows:

Furthermore, [Cooking diaries can be shared between friends via the WMF app].

One can reformulate the sentence as follows after the completeness rule in such a way that the grammar remains correct. I.e. here shell word "des Weiteren" should be integrated with the argumentation component.

Rule: If an argumentation component can be rearranged according to the completeness rule and the context of the component is not changed by this, shell words or phrases should not be included in the annotation. If the context of the argument component is changed by omitting shell words, they should be included in the annotation. Additionally, shell words in English, as well as in German, serve as indicators for identifying argument components.

#### A.2.4.4 Separation rule

Rule: A sentence may be fully annotated only if the corresponding sentence does not

contain conclusion and shell words between different opinions. That is, a sentence may not be split into two argument components if there are several complete opinions. Only if one opinion is the reason for another opinion, the sentence may be split into two or more argumentation components.

It is important that sentences containing multiple opinions are not separated if they are connected by words such as "and", "or", and bullets, and do not contain an intervening conclusion. This can occur when one sentence contains multiple reasons for an opinion from another sentence. In this case, the various reasons (premises) from one sentence must be annotated as a single argument component. This also applies to conditional sentences, which contain a condition under which something occurs (Stab and Gurevych, 2017). An indicator of such sentences are words like "if", or "in that". The subordinate clause formulates the condition that must be met in order for the consequence stated in the main clause to be realized.

If this brand has a sale, the desired garment will be automatically displayed and, if desired, delivered to your home.

P1: the desired garment will be displayed automatically AND P2: delivered to your home if desired.

In this example, the argument component contains two premises: These are connected by the word "and". Here, these two premises should be combined into one argumentation component.

#### A.2.4.5 Punctuation rule

Rule: Punctuation marks that appear at the end of an argument component must not be included in the annotation (Stab and Gurevych, 2017).

Example: [Everyone knows the problem: You have arranged to have dinner with friends, but you have no idea what to cook. In addition, one friend has an allergy to nuts and the other colleague is a vegetarian].

#### A.2.4.6 Grammar rule

Sentences that are structured like bullet points, in which grammatically important words are omitted but still represent a clear statement or premise, should be treated as if these

words were present. Likewise, sentences that are grammatically incorrect should be annotated as argument components. Statements or premises that are split into several sentences and only together make a meaningful argument component are annotated as one argument component.

### A.2.5 Quality of argumentation components

#### A.2.5.1 Step 1: Evaluation of the individual arguments and the levels

In the following, the quality of the argumentation components will be evaluated. The aim is to determine the persuasiveness of the individual argumentation components in an business pitch. The individual components (levels) are to be evaluated using the following tables, followed in the second part by an evaluation of the overall impression of the pitch. The persuasiveness of the arguments is composed of three attributes. The component types (Major Claims, Claims, and Premises) refer to the attributes of Specificity, Evidence (strength of evidence), and Strength. Specificity refers to the limited nature of a Major Claim. More specific Major Claims are more credible because they indicate the author's confidence and depth of knowledge about a topic. Evidence describes how well the supporting components describe the component being supported (Carlile et al., 2018). The attributes are each given their own score.

Specificity indicates how detailed and specific an argument component is, i.e., how characteristic or indicative the Major Claim is. Major Claims that are specific are more credible because they indicate the author's confidence and deep knowledge of a topic. The statement of a claim can be neutral, only the supporting evidence (premises) evaluate and then lead to the credibility of the claim (Carlile et al., 2018).

The aim is to evaluate how well an individual statement contributes to persuasiveness. Only premises and their influence on the claims are evaluated. The premise type takes on one of the following: User-centered example, explanation, description, invented instance, analogy, or statistic (see Carlile et al. (2018), p. 625-626). A user-centered example describes a typical single instance that is used as an explanation to help the user understand usability. The explanation attempts to make the claim understandable through logical and causal relationships. When annotating, the question should be why something is done or should be done this way, so it is about context. Questions might be "Why do I need something?" or "Why does something need to be done this way?". In contrast, description serves to convey a concrete or abstract object, state, context or circumstance to a person by linguistic means in such a way that the addressee gains a precise idea of

it. Thus, in contrast to explanation, it is a matter of mere description. When annotating, the question should be asked "How does something look like" or "How should something be done". The premise is called an invented instance when it describes a hypothetical situation. An analogy attempts to explain a claim by explaining known, similar, or partially identical circumstances. In simple terms, an analogy looks at an explanation from a different angle. With the use of statistics, numbers and data are used to justify a claim.

#### A.2.5.2 Step 2: Evaluation of the entire pitch

After the individual arguments and their components have been evaluated, the entire pitch is now to be subjected to an evaluation. The strength, clarity, and interest generation are to be evaluated. The strength of the pitch score indicates how well the idea is argued. The following score is relatively subjective compared to the scores of the first step. Therefore, they are only meant to complete the evaluation of the pitch.

**Thiemo Wambsganss - CV**

thiemo-wambsganss@gmx.de
thiemowa.github.io

## Academic Positions

04/2022 - ***Post-Doctoral Researcher*, Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, CH**
Research group head at the Machine Learning for Education Laboratory of Prof. Dr. Tanja Käser.

04/2021 – 03/2022 ***Visiting Research Collaborator*, Carnegie Mellon University (CMU), Pittsburgh, US**
Human-Computer-Interaction Institute, School of Computer Science, at the lab of Prof. Kenneth Koedinger, PhD.

04/2021 – 03/2022 ***Research Visitor,* University of Cambridge (CAM), Cambridge, UK**
Natural Language and Information Processing Research Group, Department of Computer Science and Technology, at the lab of Prof. Paula Buttery.

10/2018 – 03/2021 ***Research Assistant*, University of St. Gallen (HSG), St. Gallen, CH**
Institute of Information Management, School of Management, at the lab of Prof. Dr. Jan Marco Leimeister.
Developing end-to-end NLP-based prototypes including user studies, corpus creation, natural language modelling, interaction design, prototype deployment, and quantitative as well as qualitative evaluation studies.

## Education

10/2018 – 03/2022 **University of St. Gallen (HSG), St. Gallen, CH**
Ph.D. Program in Management (Business Innovation), Dr. oec. HSG
*Supervisor:* Prof. Dr. Jan Marco Leimeister, Institute of Information Management, University of St. Gallen
Co-Supervisor: Prof. Dr. Matthias Söllner, Information Systems and Systems Engineering, University of Kassel
Title of dissertation: *Improving Adaptive Argumentation Leaning through Artificial Intelligence*

04/2016 - 03/2018 **Karlsruhe Institute of Technology (KIT), Karlsruhe, GER**
Master of Science in Industrial Engineering, focus on Computer Science
Focus: Design Thinking, Data Analytics, Machine Learning

Title of master thesis: *Machine Learning for extracting User-Generated repair instructions from automotive web communities,* grade*:* 1.0/1

10/2012 - 04/2016 **Karlsruhe Institute of Technology (KIT), Karlsruhe, GER**
Bachelor of Science in Industrial Engineering, focus on Computer Science GPA: 1.8/1

08/2015 - 12/2015 **Beijing Institute of Technology (BIT), Beijing, CHN**
Bachelor thesis at the Institute for Software Development: *Design, implementation and evaluation of an application for a user friendly and legally compliant selection of privacy settings*, grade: 1.0/1

**Academic Awards**

2021 **Nominee Best Theory Paper Award** at **International Conference on Information Systems** for the paper "*Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education*", Austin, Texas

2021 **Best Paper Honourable Mention Award** of the **ACM *Conference on Human Factors in Computing System (CHI)*** for the paper "*ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills*", Yokohama, JP

2021 **1st place DELINA LEARNTEC Innovation Award** for the project *"ArgueLearn" - Adaptive learning of argumentation skills based on artificial intelligence, more information at https://www.learntec.de/delina/*

2021 Supervisor of **the best Master's thesis in the field of educational technology** awarded by the Gesellschaft für Informatik (German Society)

2021 **One-year research fellowship from the Swiss National Science Foundation (SNSF) -** project 200207, Pittsburgh, USA

2021 **Doctoral Consortium** at the ***ACM Conference on Human Factors in Computing System (CHI)*** *with the contribution " Designing Adaptive Argumentation Learning Systems Based on Artificial Intelligence", Yokohama, JP*

2021 **Best Reviewer Award** at the International Conference on Wirtschaftsinformatik, Essen, GER

| | |
|---|---|
| 2021 | **Nominee Best Paper Award** at the International Conference on Wirtschaftsinformatik for the paper *"ArgueBot: A Conversational Agent for Adaptive Argumentation Feedback", Essen, GER* |
| 2020 | **Best Paper Honourable Mention Award** of the **ACM *Conference on Human Factors in Computing System (CHI)*** for the paper *"AL: An Adaptive Learning Support System for Argumentation Skills", Honolulu, USA* |
| 2020 | **Best Reviewer Award** at the International Conference on Information Systems, Hyderabad, India |
| 2020 | **Best Theory Paper First Runner-Up Award** at the International Conference on Information Systems for the paper "*The Anatomy of User Experience with Conversational Agents: A Taxonomy and Propositions of Service Clues", Hyderabad, India* |

**Invited Talks**

| | |
|---|---|
| May 2022 | **Invited panelist at "The First Workshop on Intelligent and Interactive Writign Assistants"**, ACL 2022, Dublin, IR |
| Mar 2022 | **University of Cambridge, UK,** Natural Language and Information Processing Research Group hosted by Paula Buttery |
| Feb 2022 | **Data Science in Education Meetup London, UK,** hosted by Aigner Picou |
| Jan 2022 | **German Research Center for Artificial Intelligence, GER,** Educational Technology Lab hosted by Niels Pinkwart |
| Jan 2022 | **Ludwig Maximilian University of Munich, GER,** Human Centered Ubiquitous Media hosted by *Albrecht Schmidt* |
| Jan 2022 | **University of Koblenz and Landau, GER,** Institute for Education in Childhood and Adolescence hosted by Miriam Leuchter |
| Jan 2022 | **Karlsruhe Institute of Technology, GER,** Institute of Information Systems and Marketing hosted by Alexander Mädche |
| Dec 2021 | **Carnegie Mellon University, US,** Language Technologies Institute hosted by Carolyn Rose |
| Nov 2021 | **Swiss Federal Institute of Technology in Lausanne (EPFL), CH,** Machine Learning for Education Laboratory hosted by *Tanja Käser* |
| Nov 2021 | **University of Michigan**, **US**, Interactive Systems Lab and AI Lab hosted by *Xu Wang and Lu Wang* |

| | |
|---|---|
| Oct 2021 | **Carnegie Mellon University, US**, Human-Computer Interaction Institute (HCII) hosted by *Ken Koedinger* |

**Reviewing**

| | |
|---|---|
| 2022 | CSCW, Computers & Education |
| 2021 | ACM CHI, CSCW, ICIS, ECIS, HICSS, WI, Journal of Information Technology |
| 2020 | |
| 2019 | ICIS, ECIS, HICSS, WI, European Journal of International Management |
| | ICIS, ECIS, WI |

**Community Outreach**

| | |
|---|---|
| 2022 | **ACM Conference on Human Factors in Computing Systems** (CHI), **Associate Chairs** *Subcommittee Learning, Education, and Families Families* |
| 2022 | **International Conference on Information Systems (ICIS), Associate Editor** *Digital Learning and IS Curricula Track* |
| 2019 | **International Conference of Information System,** Review Volunteer |

**Professional Experience**

| | |
|---|---|
| 08/2017 - 01/2018 | **System Vertrieb Alexander GmbH, Wiesbaden, GER**<br>M.Sc. thesis about a use case of a large German automotive supplier in the field of text mining/ machine learning |
| 09/2016 - 07/2017 | **Design Thinking Project ME310, Stanford University, Stanford, CA**<br>Implementing a service innovation for a partner company with Design Thinking within the Stanford-Sugar-Network |
| 04/2016 - 07/2017 | **Karlsruhe Service Research Institute (KSRI), Karlsruhe, GER**<br>Student research assistant for the development of a methodical construction kit for the design of Big Data services |
| 09/2016 - 10/2016 | **Jaaga Startup, Bangalore, IN**<br>Designed and implemented a digital workshop concept for the startup |
| 04/2016 - 08/2016 | **Research Center for Computer Science (FZI), Karlsruhe, GER**<br>Developed a customer-oriented and cross-platform app for a large German automotive supplier |