Improving Explainability and Accuracy through Feature Engineering: A Taxonomy of Features in NLP-based Machine Learning

Completed Research Paper

Thiemo Wambsganss
University of St.Gallen
St.Gallen, Switzerland
Carnegie Mellon University
Pittsburgh, USA
thiemo.wambsganss@unisg.ch

Christian Engel
University of St.Gallen
St.Gallen, Switzerland
christian.engel@unisg.ch

Hansjörg Fromm
Karlsruhe Institute of Technology
Karlsruhe, Germany
hansjoerg.fromm@kit.edu

Abstract

Natural Language Processing (NLP)-based machine learning receives continuous attention in Information System (IS) research and practice. Despite the success of deep learning models, NLP feature engineering still plays a vital role in contexts where only little annotated data is available, and in which explainability is a precondition for productive deployment. However, NLP feature engineering is a labor-intensive and time-consuming endeavor, and there is still limited shared knowledge about the distinctive characteristics of NLP features from an interdisciplinary perspective. To address this gap, we draw on a systematic literature review and develop a five-dimensional NLP feature taxonomy based on 133 unique features from 211 scientific studies. This helps IS researchers and practitioners to classify, compare, and evaluate their NLP studies. Moreover, we used cluster heat mapping analysis to derive three clusters and several white spots to provide further assistance for designing new NLP solutions in IS.

Keywords: Natural Language Processing, Text Features, Clustering, Taxonomy

Introduction

Natural Language Processing (NLP) is receiving continued attention from both Information Systems (IS) researchers and practitioners. In the last years, several application areas have become popular, such as spam detection (Wood 2016), sports performance prediction (Gruettner et al., 2020; 2021), web mining for predicting pandemic outbreaks (Jahanbin and Rahmanian 2020), predictive policing (Meijer and Wessels 2019), conversational agents (Zierau et al. 2020) or procedural knowledge extraction (Wambsganss and Engel 2021). All of these information systems have been heavily fueled by the steep technological developments in the fields of big data, artificial neural networks (ANN), and transfer learning algorithms such as BERT (Devlin et al. 2018). The latter have proven to be very successful in processing large amounts of data and predicting outcomes precisely (Alom et al. 2019). However, in contexts, in which only little annotated data is available, and in which the explainability (London 2019), the legal accountability (Bibal et al. 2020) or the auditing of algorithms (Rahwan et al. 2019) plays a vital role for successful deployment in practice, the deployment of deep learning approaches is limited (Arrieta et al. 2019; Rahwan et al. 2019;

Stoica et al. 2017). For instance, these applications can be found in areas where processes underly high audit and regulatory requirements, such as finance and banking (Bibal et al. 2020). Often models that exhibit a low degree of explainability cannot be deployed in production due to missing transparency of what happens between data inputs and algorithmic outputs (Miller 2019). Other examples can be found in contexts where by nature of the domain, lower volumes of data are available such as for highly specific use cases, for underrepresented languages (where pre-trained models are not available) or for small data sets (Arrieta et al. 2019). In this realm, one striking example that exhibits high societal impact can be found in the domain of small and medium enterprises (SMEs). Here, the naturally lower number of data points due to limited transaction volumes restricts them from leveraging deep learning and thus prevents the backbone of our industry from seizing the full potential of digitalization for gaining competitive advantage, or to merely survive. This becomes especially relevant against the backdrop of the COVID-19 pandemic as SMEs have gotten under increased pressure to assure business continuity. This shows that there are plenty of IS use cases, in which machine learning (ML) models that incorporate "traditional" feature engineering (the combination of human domain expertise and ML) are still the more promising – and often the only feasible – modelling approach to account for explainability or lower volumes of data. One common approach for addressing explainability in large data sets, as well as for training accurate results on small amounts of rare text data, is to pre-process written text and extract valuable features (Rajman and Vesely, 2004; Johnson et al. 2015; Allahyari et al. 2017). "Feature engineering" is about enhancing NLP models with the appropriate design, implementation, and evaluation of text features to improve prediction outcomes (e.g., Bird et al. 2009; Nassirtoussi et al., 2014; Johnson et al., 2015; Fromm et al. 2019; Ribeiro et al. 2020). Recently, a text-feature perspective has also gained popularity in the field of explainability for post-hoc *"feature relevance analysis"* in order to increase the transparency of deep learning models (Arrieta et al. 2019). Here, feature engineering is used to better understand which text feature leads to which model outcome.

However, NLP feature engineering comes with several 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). Especially for SMEs, which often do not have experienced data science teams, feature engineering is a major competitive disadvantage, since in practice, "applied machine learning' is basically [only about] feature engineering" (Ng, 2013). Third, as Talib et al. (2016) point out, NLP has been influenced by different disciplines such as *Computer Science*, *Statistics*, *Computer Linguistics*, and *Information Systems*. 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 (HCI), 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 are rather scarce (Fromm et al. 2019). A few approaches that try to fill this gap already exist in literature, such as Fromm et al. (2019) or 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, it is our goal to support both practitioners and researchers in their efforts of conducting feature engineering by providing them with a conceptually sound and empirically grounded NLP feature design space. Therefore, in this paper, we develop a comprehensive multidisciplinary taxonomy of text features that shall serve as a guidance for practitioners and researchers by revealing the complex relationships of text features and by illustrating the application of certain feature groups based on similarity analysis. Thereby, we aim to contribute to a deeper understanding of NLP feature engineering by answering the following research questions (**RQs**):

RQ1: *What are the theoretically grounded and empirically validated dimensions and characteristics of NLP features?*

RQ2: *Can we observe patterns of usage of these features across research publications that help us to understand relationships between features and feature groups?*

To answer these research questions, we first illustrate the diversity of text feature engineering through a systematic literature review (vom Brocke et al. 2015), in which we found 133 distinct NLP features based on 211 publications in different applications domains. Second, we outline a classification framework through a taxonomy with five dimensions and multiple characteristics (Nickerson et al. 2013). Third, we evaluate and revise the taxonomy with five experts, and identify three clusters and several "white spots" based on a cluster analysis to provide further assistance for the development and design of new NLP solutions in IS and the assessment and comparison of existing IS applications. For practical guidance, we, present our taxonomy as an NLP feature canvas, in which a single text feature can be categorized and compared along five dimensions. Moreover, we illustrate all 133 identified text features on a web platform and encourage researchers to add missing features accordingly[1]. Our ultimate goal is to present an eclectic collection of text features and a corresponding classification framework towards a shared understanding across application domains. With that we hope to contribute to both, scientific literature by empirically analyzing the manifold use of NLP features, and to practice, by proposing a design space of features and categories incorporated in an easily usable feature canvas that shall help to approach NLP endeavors in a structured manner. The resulting taxonomy should simplify the comparability of NLP features between different studies, domains, or applications and reduce the costly effort for feature engineering of practitioners and scientists, e.g., for increasing the explainability of large data models or to improve the accuracy of small specific data sets. We present a taxonomy of features diverse enough to demonstrate their commonalities and differences and comprehensive enough to encompass every feature we found in our systematic literature review. We believe that this diversity will encourage researchers and practitioners in IS to increase the quality of their feature generation for research purposes or practical applications. This shall make ML more accessible and feasible also in context with high explainability requirements or small data sets, such as in the case of SMEs.

Theoretical Background

NLP Feature Engineering for Information Systems

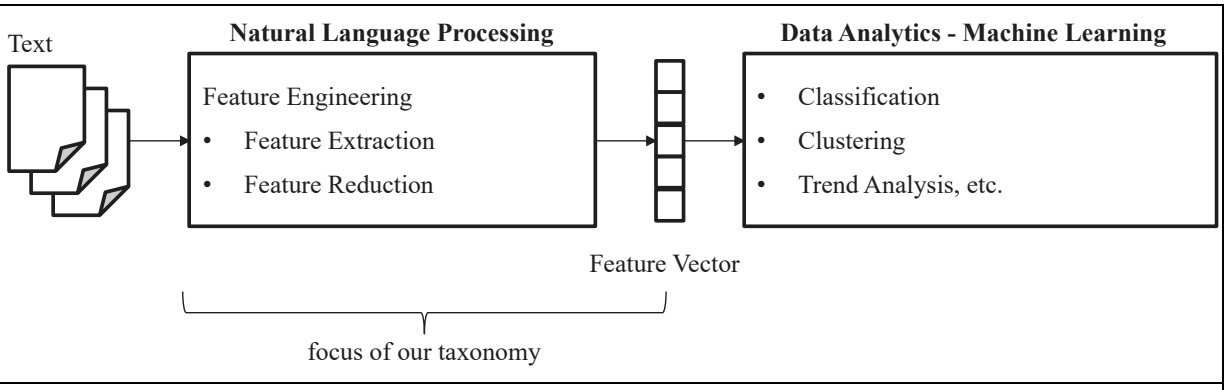

Figure 1. Illustration of pre-processing, feature engineering and NLP in an intelligent IS

Text-based Artificial Intelligence for IS is based on techniques from different areas: NLP, Data Analytics (DA), and ML. Hereby, the main challenges are to extract and reduce relevant features in a feature vector to reach a maximum performance by different DA or ML algorithms (e.g., classification, clustering, or trend analysis, etc.) as depicted in Figure 1 (Rajman and Vesely 2004). Recently, a text-feature perspective has also gained popularity in the field of explainability for post-hoc *"feature relevance analysis"* to increase the

[1] https://thiemowa.github.io/NLPFeatures/

transparency of deep learning models (Arrieta et al. 2019). Feature analysis is used to better understand which text features yield a certain predictive outcome. Features are understood as certain text characteristics or distinct attributes of a text that might bear valuable information for the ML or DA algorithms. Thus, we use NLP features and text features as synonyms. In our research, we aim to address the NLP pipeline of an intelligent text-based information system, not the modelling (e.g., the ML) pipeline (Figure 1). Fueled with large amounts of data, artificial neural networks (ANNs) 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 (e.g., Alom et al., 2019; Wambsganss, Molyndris and Söllner, 2020; Wambsganss, Söllner and Leimeister, 2020). These models are said to learn features automatically in their top layers and adjust to the patterns of the data in their lower layers. However, big amounts of data are usually necessary for those algorithms to be trained to detect and predict these data patterns. Moreover, deep ANNs *"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 ANNs (Rahwan et al. 2019; Stoica et al. 2017). 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 (Alom et al. 2019; Stoica et al. 2017). On the other hand, traditional ML approaches often show better performance for small amounts of input data (e.g., less than 1,000 annotated documents) (Alom et al. 2019). Traditional ML models together with text feature engineering are therefore still mainly used as the default modelling approach for NLP since often only small data sets are available that are annotated, e.g., for human-centered writing support systems (Wambsganss et al., 2020) 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 is creating relevant features and deciding how to represent them. Bird et al., (2009, p. 224) mention that it is possible to receive 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 taxonomy or framework. In fact, we did not find a 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., Indurkhya and Damerau (2010); Johnson et al. (2015) or Missen et al. (2013). Most of the time, feature engineering is done by intuition or expertise 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 several 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 the vocabulary richness or the use of hapax legomena (e.g., NagaPrasad et al. 2015). Procedure or instruction mining tries to find answers for "how to" questions in the vastness of the internet (Wambsganss and Fromm 2019). This can range from instructions on how to repair an automobile to recipes on how to prepare a meal. Typical features used in this domain are the occurrence of enumerations, imperatives, or certain verb-noun combinations. 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.

Taxonomies as a Means for Classifying NLP Features in IS

To better understand the relevance of a taxonomy for NLP features for research and practice, we 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 we 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, built, and deploy a new NLP application. In this sense, the ability to classify NLP text features is a prerequisite to build, compare, and assess 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. Landolt et al. (2021) offers a taxonomy for DL application in NLP. In addition, Fromm et al. (2019) published a first feature taxonomy for text mining features, however, their research was not based on a systematic literature review and did not incorporate certain feature levels. Therefore, we 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.

Research Methodology

Our 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 1). Second, we derived dimensions and characteristics in a novel taxonomy following the method presented by Nickerson et al. (2013). Third, we evaluated the quality of the taxonomy with five semi-structured interviews with experts from research and industry. Finally, we 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).

	Step 1: Database Creation	**Step 2: Taxonomy Development**	**Step 3: Taxonomy Evaluation**	**Step 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 (vom 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 phase 1)
Results	Database with 211 articles on NLP feature engineering	Taxonomy of NLP features	Evaluated taxonomy of NLP features	Insights of NLP usage patterns

Table 1. Overview of the four consecutive research steps.

Step 1: Database Creation Through a Systematic Literature Review

To identify the diversity of different text features used for feature engineering, we have drawn on the systematic literature search approaches by Webster and Watson (2002) and 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), and Johnson et al., (2015), we identified different keywords, which researchers used to describe the process of NLP feature engineering. Based on these, we 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, we 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. Table 2 summarizes the obtained hits and the relevant papers of each database.

	Databases									
	AISeL		**EBSCO**		**ScienceDirect**		**ProQuest**		**ACM Digital Library**	
	Hits	Relevant	Hits	Relevant	Hits	Relevant	Hits	Relevant	Hits	Relevant
Search String	5375	9	1099	69	2034	57	798	48	392	58
Total hits	9698									
Total relevant papers	241									
Without duplicates	206									
+ 17 forward-backward search	**223**									
-12 papers without text features	**211**									

Table 2. Overview of found hits and relevant papers for each database.

Paper selection: The database search resulted in 9698 hits. Titles, abstracts, and keywords were screened to fit the abovementioned definition and application to the scope of our study. We excluded papers that did not refer to textual data or that applied feature engineering in another domain than natural language processing. Multiple papers were excluded due to a different research scope described in their abstract, e.g., several papers described feature extraction for text recognition in images and thus were eliminated from our 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, we thoroughly read each paper. We found 12 papers that described a process or applied a method within NLP without mentioning or describing any feature. Therefore, we excluded those papers from our 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, we developed a list of master codes and master code descriptions representing different NLP features. This process was iterative and required multiple rounds of coding of the identified papers by different researchers. The process started by 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 our list until all papers were coded. Our final feature list consisted of 133 unique NLP features from 211 coded papers.

Step 2: Taxonomy Development

We aim to provide framing through the development of a comprehensive taxonomy. Therefore, we follow the method presented by Nickerson et al. (2013), which has been applied by several other studies in the IS field, such as Feine et al. (2019). The method follows an iterative and structured process for developing taxonomies grounded on theoretical foundations (deduction) and empirical evidence (induction). By applying the method of Nickerson et al., (2013), we develop different dimensions and characteristics based on both, published studies about feature engineering for NLP, found in our literature review, and empirical evidence of specific meta attributes. The development of a taxonomy usually starts with defining a specific phenomenon of interest, also called meta-characteristic. The creation of all dimensions and characteristics should be based on contributing to this meta-characteristic. Our meta-characteristic is to develop a novel artifact which facilities NLP feature engineering of scientists and practitioners by forming theoretically grounded and empirically validated dimensions and characteristics of text features in NLP. 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. We defined the following ending conditions (EC) to determine when to terminate the iterative process.

A) At least one object (text feature) is classified under every characteristic of every dimension.

B) No new dimension or characteristic has been added in the last iteration.

C) Dimensions and characteristics are unique and are not repeated.

D) Every known object (text feature) is classified in the taxonomy.

All ending conditions should be fulfilled by the final taxonomy. In the presented taxonomy below, all ending conditions are met and we believe that to the best of our knowledge we represent all the found NLP features. However, especially condition *D)* might give us difficulty in the future since we would never claim that we found all NLP features used by practitioners or scientists. However, we believe that our current taxonomy presents the diversity of text features and completeness to all identified features in the systematic literature review. We have found a set of features diverse enough to demonstrate their commonalities and differences and comprehensive enough to display all identified text features.

Iteration	Approach	Taxonomy	EC met
1	conceptual-to-empirical	T_1 = {Linguistic Category (morphological, lexical, syntactical, semantical), Granularity Level (character, word, sentence, document)}	A,C
2	empirical-to-conceptual	T_2 = {Linguistic Category (morphological, lexical, syntactical, semantical), Granularity Level (character, word, sentence, document), Information Source (corpus-based, lexicon-based)}, Dimensionality (one-dimensional, multi-dimensional), Representation (integer, real number)	A,C
3	empirical-to-conceptual	T_3= {Linguistic Category (morphological, lexical, syntactical, semantical), Granularity Level (character, word, sentence, document, beyond document), Information Source (corpus-based, lexicon-based)}, Dimensionality (one-dimensional, multi-dimensional), Representation (binary, integer, real number)	A,C
4	empirical-to-conceptual	T_4= {Linguistic Category (non-linguistic, morphological, lexical, syntactical, semantical), Granularity Level (character, word, sentence, document, beyond document), Information Source (corpus-based, lexicon-based)}, Dimensionality (one-dimensional, multi-dimensional), Representation (binary, integer, real number)	A,C
5	empirical-to-conceptual	T_5= {Linguistic Category (non-linguistic, morphological, lexical, syntactical, semantical, non-linguistic), Granularity Level (character, word, sentence, document, beyond document), Information Source (corpus-based, lexicon-based)}, Dimensionality (one-dimensional, multi-dimensional), Representation (binary, integer, real number)	A,B,C,D

Table 3. Taxonomy development iterations based on Nickerson et al. (2013).

Since we would estimate a high level of knowledge in the research area and multiple NLP studies with feature engineering are available, we conducted a conceptual-to-empirical cycle first, followed by four

empirical-to-conceptual cycles. We inductively challenged the latest status of the taxonomy by classifying NLP features and revising existing dimensions and characteristics accordingly. The development of our taxonomy is illustrated in Table 3. In total, we 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.

Step 3: Taxonomy Evaluation

To ensure the quality of our taxonomy, we assessed it against the following five criteria: *conciseness, robustness, comprehensibility, extendibility*, and *explanatory power* (Nickerson et al. 2013). Hence, to evaluate the taxonomy, we conducted semi-structured interviews with five experts that either had expertise in NLP research (four interviewees) 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 our taxonomy, the meta-characteristic, as well as exemplary NLP feature were sent to the interviewees before the interviews. In the interviews, we 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 were 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. We considered this in our 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. We 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 3).

Step 4: Taxonomy Application through Cluster Heat Mapping

Based on our finally evaluated taxonomy, we 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, we argue there may exist "natural" clusters of NLP features, which would help us to display the complex relationships of text features and better interpret their application to provide researchers and practitioners further support in their practical work. We 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, we created a data matrix $D(i,j)$, $1 \le i \le N$, $1 \le j \le M$ 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 our 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.

We 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. We 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 our 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. We believe, by visualizing the results, that we can provide further insights by identifying a natural grouping of publications that used similar features and derive possible white spots.

Taxonomy of NLP Features

In total, we found 133 distinct NLP features in 211 papers that discussed feature engineering for NLP and described at least one textual feature. We 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 we 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).

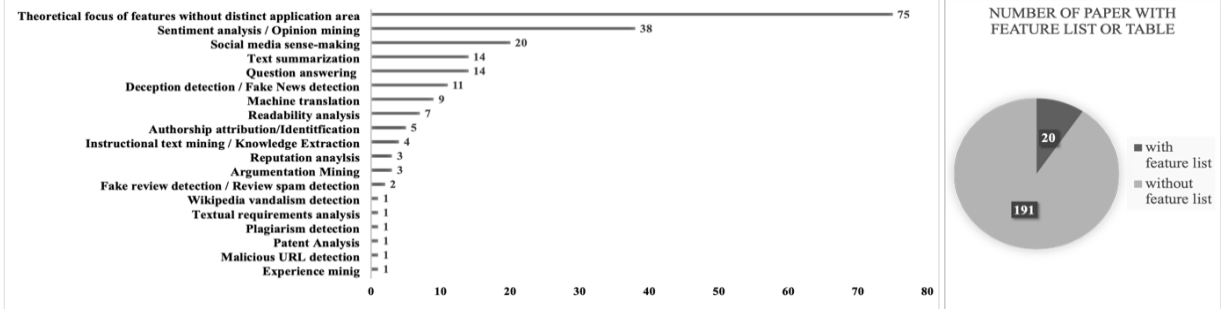

Figure 2. Overview on analyzed NLP features across domains.

In total, we found 20 papers containing a feature list or table and 191 papers which simply present their distinct features without further illustration or categorization (see Figure 2 on the right). The majority of papers that we 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 2 shows their distribution on the left. In total, we found more than 20 application areas in our 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, Mandarin, or Swedish documents.

Linguistic Analysis Level

The dimensions most widely mentioned in literature are the linguistic perspective (Johnson et al., 2015), or in other words, the hierarchy of stages in NLP: *morphological analysis, lexical analysis, syntactical analysis*, and *semantic analysis* (Bird et al., 2009; Johnson et al., 2015). 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 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. 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 our analysis, we always assign a feature to the highest linguistic

analysis level. However, some features 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). We addressed those features with a category called *non-linguistic features*.

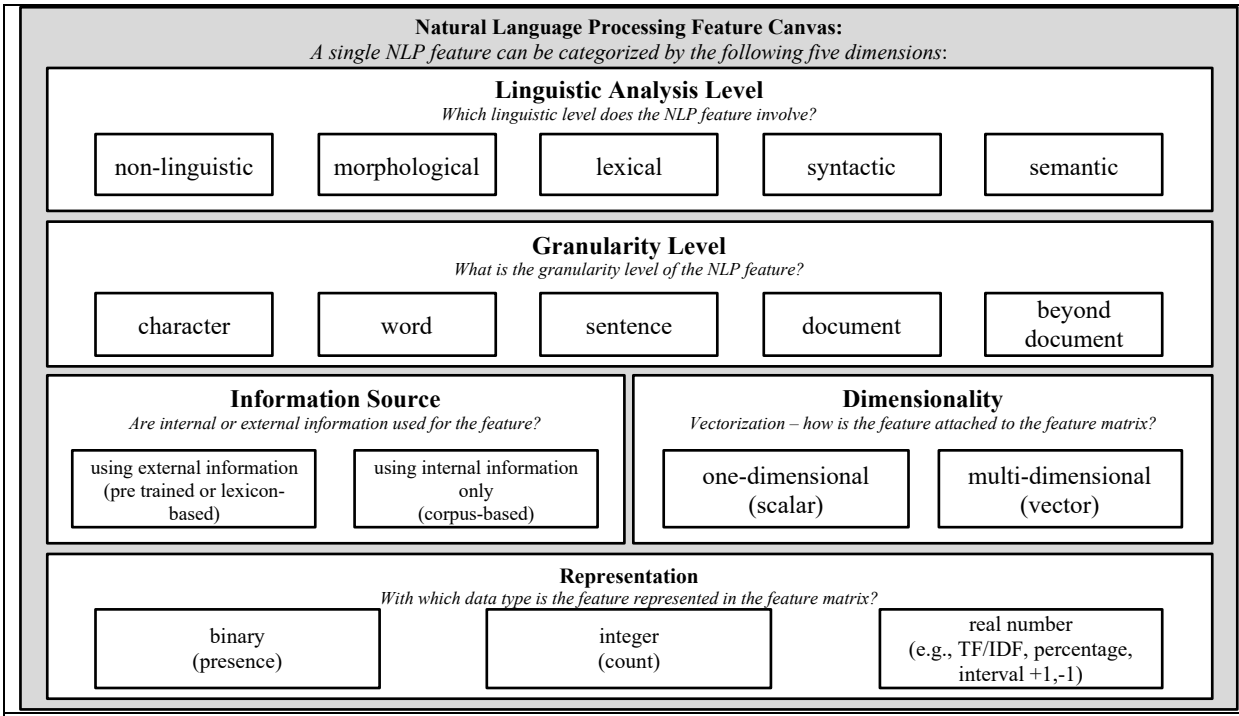

Figure 3. Taxonomy of NLP features in an NLP feature canvas according to the method proposed by Nickerson et al. (2013).

Granularity Level

A much less observed characteristic of textual features is their granularity level (so called by Missen et al., 2013 and Suh, 2016). A hierarchy can be described that decomposes a text (a document, a review, a post, a tweet) 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 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 single words (= unigrams), which have syntactic roles (= part-of-speech tags, POS tags). Occurrences of both unigrams and POS unigrams can be counted and thus be used as features. 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, we 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 non-linguistic 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 tweets 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. We address the granularity level of those features with a characteristic called *"beyond document"*.

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). WordNet 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, we included a dimension called *"information source"* with the characteristic's corpus-based and lexicon-based to provide differentiation.

Feature	References	Linguistic Analysis Level	Granularity Level	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); Tang and Cao (2015)	syntactical	word	no	vector	integer (count)
Word embeddings	(Levy and Goldberg 2014)	lexical	word	no	vector	real number
Binary character trigram vectors	Lipka and Stein, (2010)	lexical	character	no	vector	binary (presence)
Readability indices (Flesch, Kincaid, etc.)	Flesch (1943); Kincaid (1975)	lexical	document	no	scalar	real number
Word polarity	Rice and Zorn (2013); Agarwal and Mittal (2016)	semantic	word	yes	scalar	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	scalar	real number (-1; +1)
Time stamp of a message	Abbasi et al. (2018)	non-linguistic	document	no	scalar	real number
Number of links pointing to a web page	Fürnkranz (1991)	non-linguistic	beyond document	no	scalar	real number

Table 4. Feature examples categorized by linguistic analysis level, granularity level, external sources dimensionality, and representation.

Dimensionality of Features

As already stated above, several classifications of NLP features along single dimensions can be found in 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., Flesch, 1943; Kincaid et al., 1975; Feng et al. 2010), and features represented as a vector, like bags of words, bags of n-grams, word embeddings, or sentence embeddings (e.g., Levy and Goldberg, 2014; Brett and Pinna, 2015; Wambsganss et al., 2020). 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).

Representation of Features

All features that we found in the literature were expressed in numbers, also categorial features (e.g., POS-tags). 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, we 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 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. The representation dimension helps to avoid confusion in this terminology.

We want to provide a deeper understanding of the five dimension and their characteristics. Therefore, we apply our taxonomy (NLP feature canvas) to different examples illustrated in Table 4.

Cluster Heatmapping of NLP Features Across Studies

To provide further insights, we aim to identify patterns of text features by illustrating the frequency of text features according to the dimensions and characteristics we found in our sample of studies (see Figure 4). In Figure 4, we illustrate the frequency distribution of NLP features we found in our subset of NLP studies. Concerning 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 rarely use features above the word or lexical level. In fact, only 18 percent of our subset of studies about feature engineering in NLP used features on a sentence, document, or beyond-document level. This depicts a lack of application and a white spot of those feature levels since these feature characteristics showed promising values in various domains (e.g., Fürnkranz, 1999).

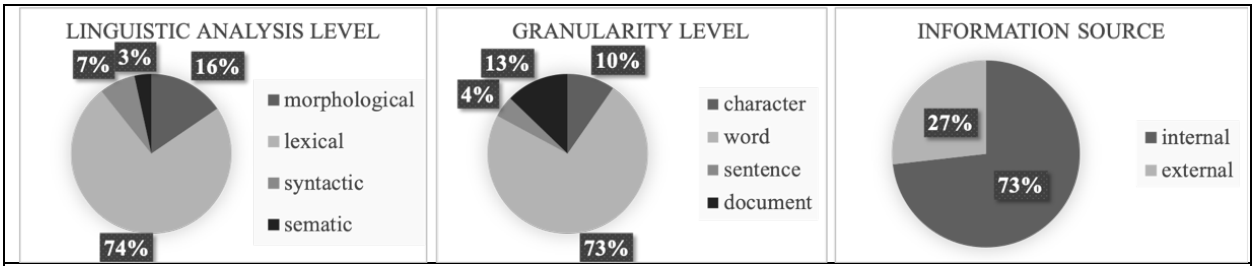

Figure 4. Illustration of the frequency distribution of text features for linguistic analysis level, granularity level, and information source in our sample of 211 NLP studies.

Next, our objective was to identify patterns and further white spots to better interpret the application of certain text feature groups that combine similar features. In Figure 5, we display our binary *NxM data matrix* before (Figure 5, left: unclustered data*)* and after rearranging the columns based on *Ward's algorithm* (Ward, 1963) (Figure 5, right: clustered data*)*. The *rows* represent the 133 NLP features, and the *columns* encompass all 211 studies from our sample collected in the systematic literature review. A black dot indicates that a feature is used in the corresponding study. Moreover, we illustrated the resulting tree in a dendrogram, the leaves of which define a similarity sequence.

In the heat map, we 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 Bag-of-Words (BoW). More than 90 percent of the analyzed papers were using BoW, which explains the long black vertical line in the heat map. Based on the dendrogram, we identified three main clusters of application patterns of studies (C1, C2, C3). Cluster one (C1 in Figure 5) represents a group of studies that solely used BoW as a feature. In fact, 50 percent of the NLP applications in our 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 BoW. 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 our taxonomy. In cluster two (C2), we see a group of studies that use BoW 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 our sample

used. Finally, we found a third cluster (C3), which reflect 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 our data set. This indicates that there seems to be untapped potential in current NLP studies.

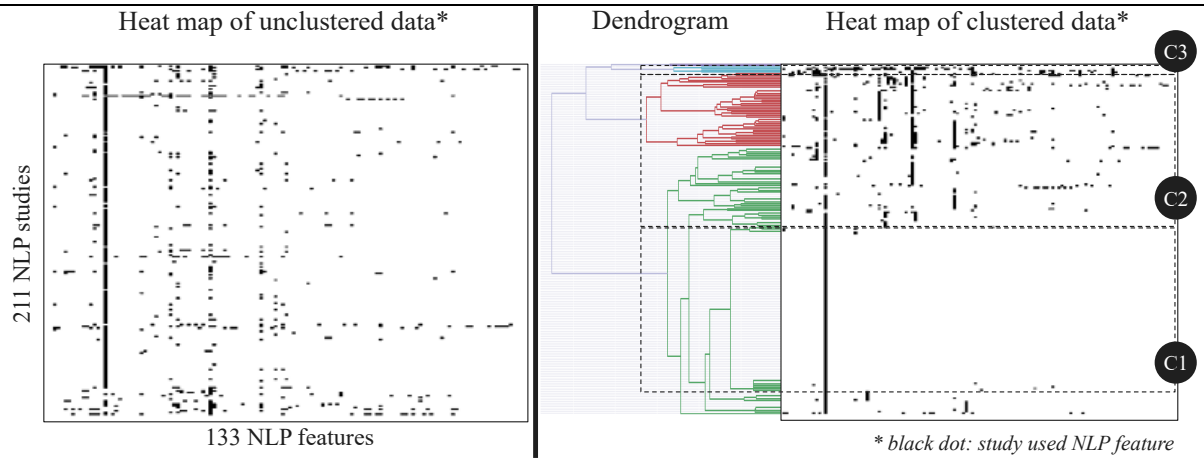

Figure 5. Illustration of clustered data matrix in a heat map (Wilkinson and Friendly 2009).

Cluster	Description	Percentage of studies	Implication
C1	Studies solely use the BoW	About 50 percent of studies in our sample	• The minority of studies tune their NLP-based ML use cases based on higher-level features (C3). • Several NLP-based ML use cases fall behind their potential (C1).
C2	Studies use BoW and additional features	About 40 percent of studies in our sample	
C3	Studies use extensive feature engineering	About 5 percent of studies in our sample	

Table 5. Overview of identified NLP feature groups (C1, C2, C3) after a hierarchical clustering.

Discussion and Contribution to Theory and Practice

We analyze 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 we found in our systematic literature review. We 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 we condensed in the taxonomy. This shall increase the quality of their feature generation for research purposes or practical applications.

From a theoretical perspective, we integrate existing literature, including literature reviews on NLP feature engineering (e.g., Fromm et al., 2019), 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, we provide a deeper understanding of what needs to be considered when designing, evaluating, and comparing NLP applications. By combining the dimensions and characteristics of our taxonomy, we 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. Our 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; Usman et al. 2017). Our

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, we 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. We believe that our 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 SMEs and other organizations to better design NLP models in their specific use cases. Moreover, our 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 application.

Given the immense growth of NLP-based ML in research and practice, as well as the interest in explainable modelling 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. In this paper, we offer such an understanding by presenting our taxonomy, its application by performing cluster analysis on it, which facilitates us to identify white spots in literature, which are worthwhile to be investigated by future research. In this regard, we offer researchers to apply a more nuanced perspective when training NLP-based models based on our 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 HCI. Our 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, our 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 our taxonomy, researchers can now identify different kinds of NLP features by combining different characteristics of our taxonomy, depending on their domain and use case.

Of course, this research has several limitations, which provide avenues for future research. First, we only identified NLP features based on scientific publications, since our 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 our taxonomy based on an in-depth analysis of real-world use cases. Moreover, in the preliminary screening, we only selected research papers that explicitly mentioned text feature engineering as our 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. We believe that with our web platform we can further contribute to increase the levels of transparency and prediction accuracy of NLP features in many domains. Finally, based on technological advancements, more NLP features may need to be added in the future. However, encouraged by the evaluation with experts, we believe that our taxonomy represents a valuable state-of-the-art tool for analyzing NLP features for current IS research and practice.

Conclusion

In this study, we developed a taxonomy of NLP features based on Nickerson et al.'s (2013) methodological approach, which allowed us to identify groups and white spots based on a cluster heat map analysis. Our NLP feature taxonomy is intended to help researchers and practitioners to develop, refine, compare, and evaluate their NLP studies and projects. In addition to this article, we freely present all 133 identified text features on a web platform and encourage researchers to add features not listed accordingly. Our ultimate goal is to present an eclectic collection of text features and a corresponding classification framework towards a shared understanding across application domains.

Acknowledgments

We thank the Swiss National Science Foundation for funding parts of this research (200207).

References

Abbasi, A., Chen, H., and Salem, A. 2008. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," ACM Transactions on Information Systems (26:3), pp. 1–34.

Abbasi, A., Zhou, Y., Deng, S., and Zhang, P. 2018. "Text Analytics to Support Sense-Making in Social Media: A Language-Action Perspective," MIS Quarterly: Management Information Systems (42:2), pp. 427–464.

Allahyari, M., Trippe, E. D., and Gutierrez, J. B. 2017. A Brief Survey of Text Mining : Classification , Clustering and Extraction Techniques.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., Asari, V. K., Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., and Asari, V. K. 2019. "A State-of-the-Art Survey on Deep Learning Theory and Architectures," Electronics (8:3), p. 292.

Alzahrani, S. M., Salim, N., Abraham, A., and Member, S. 2012. "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS (Vol. 42), p. 133.

Anaconda. 2020. "State of Data Science 2020."

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2019. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion (58), Elsevier B.V., pp. 82–115.

Baccianella, S., Esuli, A., and Sebastiani, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.

Bailey, K. D. 1994. Typologies and Taxonomies: An Introduction to Classification Techniques, CA: Sage.

Bibal, A., Lognoul, M., de Streel, A., and Frénay, B. 2020. "Legal Requirements on Explainability in Machine Learning," Artificial Intelligence and Law, Springer, pp. 1–21.

Bird, S., Klein, E., and Loper, E. 2009. Natural Language Processing with Python, O'Reilly.

Brett, D., and Pinna, A. 2015. "Patterns, Fixedness and Variability: Using PoS-Grams to Find Phraseologies in the Language of Travel Journalism," Procedia - Social and Behavioral Sciences (198:Cilc), Elsevier BV, pp. 52–57.

vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., and Cleven, A. 2009. "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," in Proceedings of the 17th European Conference on Information Systems (ECIS), Verona, Italy, 2009, pp. 2206–2217.

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., and Cleven, A. 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), pp. 205–224.

Cossu, J. V., Dugue, N., and Labatut, V. 2015. "Detecting Real-World Influence through Twitter," Proceedings - 2nd European Network Intelligence Conference, ENIC 2015, pp. 83–90.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

Doty, D. H., and Glick, W. H. 1994. "Typologies as a Unique Form of Theory Building: Toward Improved Understanding and Modeling," The Academy of Management Review (19:2), The Academy of Management, p. 230.

Feine, J., Gnewuch, U., Morana, S., and Maedche, A. 2019. "A Taxonomy of Social Cues for Conversational Agents," International Journal of Human Computer Studies (132), Academic Press, pp. 138–161.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. 2010. A Comparison of Features for Automatic Readability Assessment, pp. 276–284.

Flesch, R. 1943. "Marks of Readable Style; a Study in Adult Education.," Teachers College Contributions to Education (897).

Forbes. 2017. "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says."

Fromm, H., Wambsganss, T., and Söllner, M. 2019. "Towards a Taxonomy of Text Mining Features," in European Conference of Information Systems (ECIS), pp. 1–12.

Fürnkranz, J. 1999. Exploiting Structural Information for Text Classification on the WWW, pp. 487–497.

Gruettner, A., Vitisvorakarn, M., Wambsganss, T., Rietsche, R., and Back, A. 2020. "The New Window to Athletes' Soul-What Social Media Tells Us About Athletes' Performances," Hawaii International Conference on System Sciences (HICSS).

Gruettner, A., Wambsganss, T., and Back, A. 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), Inderscience Publishers, pp. 314–338.

Günal, S., Ergin, S., Gülmezoğlu, M. B., and Gerek, Ö. N. 2006. "On Feature Extraction for Spam E-Mail Detection," in Lecture Notes in Computer Science, Springer Verlag, pp. 635–642.

Hambrick, D. C. 1984. "Taxonomic Approach to Studying Strategy: Some Conceptual and Methodological Issues," Journal of Management (10:1), pp. 27–41.

Hatzivassiloglou, V., and Wiebe, J. M. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity, pp. 299–305.

Jahanbin, K., and Rahmanian, V. 2020. "Using Twitter and Web News Mining to Predict COVID-19 Outbreak," Asian Pacific Journal of Tropical Medicine (March), pp. 26–28.

Jeon, J., Croft, W. B., Lee, J. H., and Park, S. 2006. "A Framework to Predict the Quality of Answers with Non-Textual Features."

John, A., Premjith, P. S., and Wilscy, M. 2017. "Extractive Multi-Document Summarization Using Population-Based Multicriteria Optimization," Expert Systems with Applications (86), p. 385–397.

Johnson, K. 2007. "An Overview of Lexical Semantics," Philosophy Compass (3:1), Wiley, pp. 119–134.

Johnson, S. L., Safadi, H., and Faraj, S. 2015. "The Emergence of Online Community Leadership," Information and Organization (July), pp. 35–68.

Kambhatla, N. 2004. "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations," in Proceedings of the ACL 2004, Morristown, NJ, USA: Association for Computational Linguistics, pp. 22-es.

Kaufman, L., and Rousseeuw, P. J. 2005. Finding Groups in Data : An Introduction to Cluster Analysis, Wiley.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. 2014. "Text Mining for Market Prediction: A Systematic Review," Expert Systems with Applications (41:16), pp. 7653–7670.

Kincaid, J., Fishburne, R., Rogers, R., and Chissom, B. 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.

Landolt, S., Wambsganss, T., and Matthias, S. 2021. "A Taxonomy for Deep Learning in Natural Language Processing," in Hawaii International Conference on System Sciences (HICSS).

Levy, O., and Goldberg, Y. 2014. "Dependency-Based Word Embeddings," 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference (Vol. 2).

Liao, S., and Grishman, R. 2010. "Large Corpus-Based Semantic Feature Extraction for Pronoun Coreference."

London, A. J. 2019. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," Hastings Center Report (49:1), John Wiley and Sons Inc., pp. 15–21.

Loni, B., Van Tulder, G., Wiggers, P., Tax, D. M. J., and Loog, M. 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) (Vol. 6836 LNAI), Springer, Berlin, Heidelberg, pp. 243–250.

Mahajan, R., and Zaveri, M. 2017. SVNIT @ SemEval 2017 Task-6: Learning a Sense of Humor Using Supervised Approach, pp. 411–415.

Meijer, A., and Wessels, M. 2019. "Predictive Policing: Review of Benefits and Drawbacks," International Journal of Public Administration (42:12), Taylor and Francis Inc., pp. 1031–1039.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences," Artificial Intelligence (267), pp. 1–38.

Missen, M. M. S., Boughanem, M., and Cabanac, G. 2013. "Opinion Mining: Reviewed from Word to Document Level," Social Network Analysis and Mining (3:1), pp. 107–125.

NagaPrasad, S., Narsimha, V. B., Vijayapal Reddy, P., and Vinaya Babu, A. 2015. "Influence of Lexical, Syntactic and Structural Features and Their Combination on Authorship Attribution for Telugu Tex," in Procedia Computer Science (Vol. 48), Elsevier B.V., January 1, pp. 58–64.

Negi, S., and Buitelaar, P. 2014. INSIGHT Galway: Syntactic and Lexical Features for Aspect Based Sentiment Analysis, pp. 346–350.

Ng, A. 2013. "Machine Learning and AI via Brain Simulations."

Nickerson, R. C., Varshney, U., and Muntermann, J. 2013. "A Method for Taxonomy Development and Its Application in Information Systems," European Journal of Information Systems (22:3), Taylor & Francis, pp. 336–359.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval (2:12), pp. 1–135.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A.' Sandy,' Roberts, M. E., Shariff, A., Tenenbaum, J. B., and Wellman, M. 2019. "Machine Behaviour," Nature (568:7753), pp. 477–486.

Rajman, M., and Vesely, M. 2004. From Text to Knowledge: Document Processing and Visualization: A Text Mining Approach, Springer, Berlin, Heidelberg, pp. 7–24.

Reyes, A., and Rosso, P. 2012. "Making Objective Decisions from Subjective Data: Detecting Irony in Customer Reviews," in Decision Support Systems (Vol. 53), North-Holland, November 1, pp. 754–760.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. 2020. "Beyond Accuracy: Behavioral Testing of NLP Models with Checklist," ArXiv.

Scanlan, J. 2019. "Auditing Predictive Policing," Brigham Young University Prelaw Review (33).

Stoica, I., Song, D., Popa, R. A., Patterson, D., Mahoney, M. W., Katz, R., Joseph, A. D., Jordan, M., Hellerstein, J. M., Gonzalez, J. E., Goldberg, K., Ghodsi, A., Culler, D., and Abbeel, P. 2017. A Berkeley View of Systems Challenges for AI.

Suh, J. H. 2016. Comparing Writing Style Feature-Based Classification Methods for Estimating User Reputations in Social Media.

Szopinski, D., Schoormann, T., and Kundisch, D. 2019. "BECAUSE YOUR TAXONOMY IS WORTH IT: TOWARDS A FRAMEWORK FOR TAXONOMY EVALUATION," in Proceedings of the 27th European Conference on Information Systems (ECIS).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2011. "Lexicon-Basedmethods for Sentiment Analysis," Computational Linguistics (37:2), pp. 267–307.

Talib, R., Hanif, M. K., Ayesha, S., and Fatima, F. 2016. "Text Mining: Techniques, Applications and Issues," International Journal of Advanced Computer Science and Applications (7:11), pp. 414–418.

Usman, M., Britto, R., Börstler, J., and Mendes, E. 2017. "Taxonomies in Software Engineering: A Systematic Mapping Study and a Revised Taxonomy Development Method," Information and Software Technology, Elsevier BV, pp. 43–59.

Wambsganss, T., and Engel, C. 2021. "Using Deep Learning for Extracting User-Generated Knowledge from Web Communities," ECIS 2021 Research Papers.

Wambsganss, T., and Fromm, H. 2019. "Mining User-Generated Repair Instructions from Automotive Web Communities," in Proceedings of the 52nd Hawaii International Conference on System Sciences (Vol. 6), Hawaii, pp. 1184–1193.

Wambsganss, T., Molyndris, N., and Söllner, M. 2020. “Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach,” in 15th International Conference on Wirtschaftsinformatik, Potsdam, Germany.

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Leimeister, J. M., and Handschuh, S. 2020. “AL : An Adaptive Learning Support System for Argumentation Skills,” in ACM CHI Conference on Human Factors in Computing Systems, pp. 1–14.

Wambsganss, T., Söllner, M., and Leimeister, J. M. 2020. "Design and Evaluation of an Adaptive Dialog-Based Tutoring System for Argumentation Skills," in International Conference on Information Systems (ICIS), Hyderabad, India.

Ward, J. H. 1963. "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association (58:301), JSTOR, p. 236.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review.," MIS Quarterly (26:2), xiii–xxiii.

Weiss, S. M., Indurkhya, N., and Zhang, T. 2010. Fundamentals of Predictive Text Mining.

Wood, L. 2016. "Artificial Intelligence (AI) Market By Technology (Machine Learning, Natural Language Processing (NLP), Image Processing, And Speech Recognition), Application & Geography - Global Forecast To 2020," GII Research

Zierau, N., Wambsganss, T., Janson, A., Schöbel, S., and Leimeister, J. M. 2020. "The Anatomy of User Experience with Conversational Agents : A Taxonomy and Propositions of Service Clues," in International Conference on Information Systems (ICIS)., pp. 1–17.