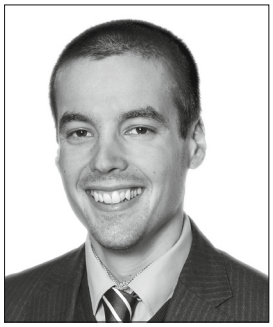
AJP/PJA 4/2022



How NLP Can Improve the Quality of Arguments in Legal Studies

A Study on the Capabilities of Natural Language Processing



Reto Gubelmann*

Peter Hongler**

In this article, we argue that there is great potential in Natural Language Processing (NLP) to qualitatively improve legal research by confronting scholars with normative points of view that are fundamentally opposed to their own by making normative backgrounds explicit that are presupposed by legal arguments. In order to do so, we will explore the specific promise of current approaches in NLP concerning normative theories in tax law. The article follows a Quinean-Pragmatist account of the epistemology of normative reasoning, emphasizing the holistic nature of disagreements regarding essentially contested concepts such as justice and the impossibility of arriving at the one true position, necessitating openness and free play of disputes.

In diesem Artikel argumentieren wir, dass Natural Language Processing (NLP) ein grosses Potenzial hat, die juristische Forschung qualitativ zu verbessern, indem es Wissenschaftler mit normativen Standpunkten konfrontiert, die ihren eigenen grundlegend entgegengesetzt sind, und normative Hintergründe explizit macht, die von juristischen Argumenten vorausgesetzt werden. Dazu werden wir das spezifische Potenzial aktueller NLP-Ansätze in Bezug auf normative Theorien im Steuerrecht untersuchen. Der Artikel folgt einer pragmatischen Darstellung der Epistemologie normativen Denkens nach Quine, die den holistischen Charakter von Meinungsverschiedenheiten in Bezug auf wesentlich umstrittene Begriffe wie Gerechtigkeit und die Unmöglichkeit, zu einer einzig wahren Position zu gelangen, hervorhebt, was Offenheit und freies Spiel von Kontroversen erforderlich macht.

Inhaltsübersicht

I. Introduction: NLP, Normative Reasoning, and Essentially Contested Concepts
II. Normative Reasoning in the Legal Discipline
III. How Do Lawyers Approach Normative Questions?
IV. Where Quine and Habermas Meet: A Pragmatist View on the Epistemology of Normative Reasoning
V. Neural NLP for Normative Reasoning
  A. General Introduction to the State of the Art in Neural NLP
  B. An Overview of the State of The Art in Natural Language Inference
  C. Exploring PLMs' Sensitivity to Normative Claims in Legal Texts
VI. How NLP Can Improve Normative Reasoning in Legal Studies: Getting Specific
VII. Conclusion

I. Introduction: NLP, Normative Reasoning, and Essentially Contested Concepts

In this article, we use the case of normative reasoning to illustrate how natural language processing (NLP) will not only lead to quantitative, but also to qualitative improvements to legal studies. It is often (and rightly) assumed that NLP will support legal researchers in the work-intensive, repetitive part of their work, say by finding and summarizing pertinent cases as well as the state of the art in research. In addition to these quantitative improvements, we argue that NLP will also lead to qualitative improvements: better research in legal studies.

We develop our argument using the case of normative reasoning. We consider it particularly suited for our argument, as normative reasoning, revolving around central normative concepts such as justice, fairness, or equality, might be taken as a field where genuinely human perspectives are indispensable. It might be held that only human beings are able to reflect on the different conceptions of justice involved in tax law, for instance. As a consequence, if we can make plausible that NLP can improve these subtle normative debates, then it seems all the more plausible that it can do so in other areas of legal research as well.

From a philosophical perspective, an important characteristic of normative reasoning is that it often revolves around so-called essentially contested concepts. This conception was first proposed by Gallie,[1] who argued that concepts such as fairness, democracy, or justice are such that essential parts of their meaning are disputed. We focus on

* Reto Gubelmann, Dr. phil., University of St. Gallen.
** Peter Hongler, Prof. Dr. iur., University of St. Gallen.

[1] Walter Bryce Gallie, Essentially contested concepts, Proceedings of the Aristotelian society 1955, 167 et seq. For recent discussions see David Collier/Fernando Daniel Hidalgo/Andra Olivia Maciuceanu, Essentially contested concepts: Debates and applications, Journal of political ideologies 2006, 211 et seq.; Philippe-André Rodriguez, Human dignity as an essentially contested concept, Cambridge Review of International Affairs 2015, 743 et seq.

essentially contested concepts or terms which reflect what is commonly known as a value-based judgment.[2]

One reason for the disputes around essentially contested concepts is that the disagreements are due to larger-scale differences in worldviews. For instance, to a Libertarian, the state has as its main, or perhaps even sole, role the protection of the individuals' life and property. Hence, any kind of taxation that surpasses the absolute minimal amount necessary to ensure that such a minimal state is functional is per se unjust. To an Egalitarian, in contrast, the state serves a much larger variety of functions. In particular, it needs to establish a maximal amount of equality among its citizens. This view of the purpose of the state justifies taxation of a much more comprehensive extent. These views of the purpose of the state, in turn, are tightly connected to substantially different views of the individual and their legal status vis-à-vis their society. Hence, the Egalitarian's and the Libertarian's conception of tax justice differ substantially, and this difference is embedded in different views on the purpose of the state, the rights and duties of individuals, the moral assessment of economic success, etc.

Our article is structured as follows. We will first discuss the importance of normative reasoning in the legal discipline and the way lawyers tend to approach this kind of reasoning (sections II. and III.), and we give a Quinean-Pragmatist account of the epistemology of normative reasoning (section IV.). Secondly, we will sketch the state of the art of NLP in general and of natural language inference in particular as well as explore the specific promise of current approaches for normative theories in tax law (section V.). Finally, we will provide arguments on why NLP could lead to said qualitative improvements (section VI.).

II. Normative Reasoning in the Legal Discipline

The goal of this article is not to develop a comprehensive understanding of the role of normative reasoning in legal research and practice.[3] However, we want to demonstrate how natural language processing can and (likely) will change the way we approach normative questions as lawyers, leading to qualitative improvements. To do so, we briefly sketch our understanding of normative reasoning.

In short, normative reasoning is relevant for the following two tasks of the legal discipline:

– *Interpretation of existing provisions:* The way in which lawyers interpret legal provisions is highly influenced by normative reasoning. Through the process of interpretation, the interpreter demonstrates how he or she understands a certain provision (i.e., how a certain provision ought to be understood from the interpreter's perspective). Such process often includes implicit or explicit reference to values such as fairness or justice, which are essentially contested. These lines of argumentation are, *prima facie*, highly influenced by personal judgments. Of course, interpretation methodology narrows the discretion, but nevertheless, value-based judgments are of key importance in legal interpretation, and they often seem to be partial. To be more precise, values are required for the following situations:[4]
  – In case of discretionary provisions, i.e., a provision that grants a judge a certain discretionary leeway.
  – In case of legal gaps. A fact pattern is not covered by the applicable law, either because the legislator has purposively decided not to regulate a certain situation, or because the legislator has not thought of a certain situation.
  – In case of concepts which require interpretation, if only because language is per se not clear and creates uncertainty.
  – In case of a collision of rules or interests. This is the case if, for instance, the application of two provisions would lead to two different results.

  In the following, we will not review classical debates about the concept of law and to what extent morality and written provisions might conflict.[5] Moreover, it is not the purpose of the present paper to discuss which legal methods should be used in what form of hierarchy.
– *Legislative developments:* Lawyers play a key role in the development of new positive rules – be it as members of Parliament or (perhaps even more importantly) as members of the executive body responsible for drafting new laws. Normative judgments are essential in the argumentation of lawyers either in favor of or

---

[2] Concerning the question what a value or value-based judgment is in the legal field, see e.g., Julia Hänni, Vom Gefühl der Rechtsfindung, Ph.D. Thesis, University of St. Gallen 2010, 6 et seq.

[3] For such a comprehensive understanding, see e.g., Joseph William Singer, Normative Methods for Lawyers, UCLA Law Review 2009, 899 et seq.

[4] These categories have partly been derived from Hänni (n. 2), 51 et seq. There are other possibilities of categorization.

[5] We will, for instance, not deal with the Hart-Dworkin Debate about the importance of legal principles and the question of what law is.

against a certain provision. Often, lawyers ground their decisions in constitutional principles or general principles of law. However, the interpretation of these principles often requires value-based judgments, and reference is again often made to essentially contested concepts such as justice or fairness. These references might often seem partial.

III. How Do Lawyers Approach Normative Questions?

As mentioned, the question of how a provision should be understood or how a provision should be designed is often answered by reference to essentially contested concepts such as justice or fairness. Such reference might be implicit. For example, lawyers might argue in the following manners:

– Interpretation A is persuasive as it leads to a just result.
– Rule A is better than rule B as it leads to a fair balance of rights and duties.

To illustrate this with a real-world example, the following argument contains an implicit reference to fairness:

> *«In general, the Supreme Administrative Court agreed with the statements made in the lower courts that the guidelines presented in the OECD Report on transfer pricing – although not binding on the Swedish tax authorities – give a favorable and* ***well-balanced elucidation*** *of the problems involved and are not contrary to Section 43, Subsection 1, of the Municipal Tax Act.» [emphasis added]*[6]

The term *«well-balanced elucidation»* contains an implicit reference to a fair result in the sense that various interests are balanced. However, the statement does not provide any details as to what a fair balance means.

Statements can also be explicit with reference to an essentially contested concept but lack any further details on how such essentially contested concept is to be understood – let alone make a case for why it shall be understood that way. For instance, statements such as the following are not uncommon in the legal discipline:

> *«That in the absence of any permanent establishment in India in terms of DTAA [Double Taxation Avoidance Agreement], no profits alleged to have accrued to the appellant are there and consequently the levy of tax as made by the AO [= tax authority] is arbitrary,* ***unjust*** *and bad in law.» [emphasis added]*[7]

Importantly, the extent to which an author deals with essentially contested concepts differs significantly. Some authors might try to review normative theories such as libertarianism or Rawls «A Theory of Justice» within a legal analysis. Other authors simply deny any reference to normative theory either by explicitly arguing that within the scope of a certain study, an in-depth analysis of the normative question is not possible, or (worse) by not even trying to answer the normative question of what just or fair means.

Therefore, it is key to analyze whether NLP can help to achieve more reflection and less arbitrariness in legal argumentation involving normative theories and essentially contested concepts. We ground our approach on a philosophy of language as well as an epistemology that has stood the test of decades of debate in philosophy, namely a general kind of pragmatism that unites the thinking of otherwise strongly different researchers such as Robert Brandom, Jürgen Habermas, and Willard Van Orman Quine, some of the most influential philosophers of the 20th century.[8]

IV. Where Quine and Habermas Meet: A Pragmatist View on the Epistemology of Normative Reasoning

In contemporary epistemology, there has been a turn towards pragmatism from various directions. While we will here sketch Willard Van Orman Quine's argumentative path to this destination, the goal, a pragmatist epistemology, is shared by a number of key figures in contemporary and 20th century philosophy, such as Jürgen Habermas or Robert Brandom.

When addressing central questions, such as the question of what constitutes a just taxation system, pragmatists do not try to deduce the *absolutely, indubitably true* position in the *solitude of* their own study from self-evident principles; rather, they trust in *common* endeavors, be it Habermas's famous ideal discourse,[9] Brandom's game

---

6 Regeringsrätten (Supreme Administrative Court), Case number 117-1989, 22.10.1991, as outlined and translated in the IBFD Case Dataset.

7 ITAT, Mumbai, Mahindra and Mahindra Ltd. vs. DCIT, ITA Nos. 2606, 2607, 2613 and 2614/Mum/2000, 9.4.2009, Asst. Year: 1998–99.

8 The latter is the second-most quoted contemporary philosopher in the authoritative Stanford Encyclopedia of Philosophy, see Eric Schwitzgebel, The Splintered Mind: The 295 Most-Cited Contemporary Authors in the Stanford Encyclopedia of Philosophy, Internet: http://schwitzsplinters.blogspot.com/2019/08/the-295-most-cited-contemporary-authors.html (last consulted on 27.8.2021).

9 See, in particular, Jürgen Habermas, Diskursethik – Notizen zu einem Begründungsprogramm, in: Jürgen Habermas (ed.), Moralbewusstsein und kommunikatives Handeln, 1st ed., Frankfurt am Main 1983; and Jürgen Habermas, Erläuterungen zur Diskursethik, Frankfurt am Main 1991.

of giving and asking for reasons,[10] or Quine's science, to result in *the best possible position*. Pragmatists typically reject the idea that the goal is to reach a final, apodictically true position. Rather, they will insist that what results from the very best of today's empirical methods, or what has passed the back and forth of intense philosophical dispute, should be accepted as justified until further argument or better evidence proves it wrong.[11]

Quine grounds his version of pragmatism in his verification holism. Quine has famously (and somewhat notoriously) rejected the idea that individual hypotheses could be tested empirically; rather, only a sufficiently inclusive portion of theory can be subjected to such a test. In *Two Dogmas of Empiricism*, Quine boldly claims that it takes «the whole of science» to arrive at a testable empirical prediction.[12] This means that, as it were, every single bit of evidence is spread across the whole of science, potentially affecting each and every statement of it. We call this position verification holism. Quine never abandoned this verification holism, though he attenuated it somewhat, insofar as he came not to consider the whole of science as the minimal empirically significant unit, but rather sufficiently inclusive portions of theory, for example in his article Five Milestones of Empiricism[13] or in his late Pursuit of Truth.[14]

If verification holism is right, then it is misguided to claim that there is a specific class of statements which are true regardless of any empirical evidence. If it is true that every piece of empirical evidence always calls into question entire chunks of theory—including (among others) logical law statements, epistemic norms, and ontological claims—then every statement whatsoever can be called into question by empirical evidence. For instance, if empirical scientists (or philosophers) decide that, given certain evidence, it is fit to abandon the law of the excluded middle,[15] the logician is unable to provide reasons against this that are solely based on logical considerations without reference to experience. Rather, she will have to show that the overall theory is more elegant and still able to accommodate the evidence if the law of the excluded middle is kept in place. Thus, the contribution to successful prediction of empirical data—supplemented with some rough and ready notions of theoretical elegance and simplicity—becomes the fundamental checkpoint for any claim whatsoever.[16]

Quine has famously declared that this leads him to embrace a more «thorough pragmatism» than the one of his teacher Rudolf Carnap.[17] Instead of seeing science as the pursuit of absolute, objective truth, Quine argued that we should be more modest: we should see science as a rational quest, but we should be aware that its results are always provisional (held true until proven wrong) and resting on decisions by humans that are not determined by scientific results but rather by considerations of efficiency or simplicity.

For the field of legal reasoning, the implications of this Quinean holism are significant; on a first approximation, they are largely congenial to other pragmatic thinkers of the 20th century, such as Robert Brandom or Jürgen Habermas. According to pragmatists, it is not by retreating to solitude, but rather by being open to experience, conceived very broadly, that the scholar might hope to improve her position; and the truth of this position can never be established absolutely—if it passes the litmus test of the back and forth of the game of giving and asking for reasons, it is held true until a better argument, a more comprehensive perspective proves otherwise.

Furthermore, as in Quine's famous web of belief (being a direct consequence of his verification holism), we should accept that our adherence to a certain conception of justice is influenced by and depends on other nodes in this web: we have illustrated this in our discussion of essentially contested concepts for the case of just taxation. Truly appreciating a conception of justice different from one's own therefore often requires questioning entire chunks of our own web rather than individual concepts.

Therefore, confrontation with other ideas, arguments or concepts in a broad scope might question our normative stance and thereby improve it.[18] In other words, it seems that normative judgments of lawyers might become

---

10 See, in particular, Robert Brandom, Making it explicit: Reasoning, representing, and discursive commitment, Cambridge (Massachusetts) 1994.

11 This is congenial to van Fraassen's seminal conception of Empiricism as a Stance, see Bas van Fraassen, The Empirical Stance, New Haven 2002.

12 Willard Van Orman Quine, Two Dogmas of Empiricism, in: From a Logical Point of View, Cambridge (Massachusetts) 1980 (1951), 42.

13 Willard Van Orman Quine, Five Milestones of Empiricism, in: Theories and Things, Cambridge (Massachusetts) 1981, 71.

14 Willard Van Orman Quine, Pursuit of Truth, 2nd ed., Cambridge (Massachusetts) 1992, 14–15.

15 I.e., that any proposition is either true or false, that there is no middle ground between these.

16 See Reto Gubelmann, A Science-Based Critique of Epistemological Naturalism, Basingstoke 2019, ch. 6.

17 Quine (n. 12), 46.

18 We are not suggesting that Quine's naturalization of epistemology will also lead to a naturalization of law or the philosophy of law (see on the latter dispute Mark Greenberg, Naturalism in Epistemology and the Philosophy of Law, Law and Philosophy 2011, 419 et seq.).

more appropriate the more they face recalcitrant evidence and arguments. It has also been argued by others that values are not «stubbornly there»,[19] but rather develop over time.[20]

Let us refer to an example in order to explain this further. The following is a normative statement often found in tax law analysis.

(N 1) Treating people equally before the law requires that people with the same amount of income shall be taxed at an equal rate.

In a classical, non-pragmatic epistemology, the truth of (N 1) could be assessed in a number of ways, for instance in a traditional rationalist manner: By reflecting *a priori* about the nature of equality, of the state, and of its right to levy taxes. This approach has a number of problems. To begin with, the history of philosophy shows that humans are not as proficient at grasping the nature or essence of things as some enlightenment rationalists have hoped: there is by no means an agreement on the truth of (N 1). More importantly, perhaps, it is doubtful in principle whether such natures or essences do exist, and if they exist, whether they are accessible in principle to human understanding. This is precisely the problem that Kant tried to address in his Copernican turn in philosophy, developing a transcendental approach that has redefined epistemology.[21]

In stark contrast, empiricists such as John Locke,[22] David Hume,[23] or John Stuart Mill would try to assess (N 1) using empirical observation, defined quite narrowly. Mill would combine this empiricist epistemology with a utilitarianist moral approach,[24] which would allow him to challenge (N 1) by empirically assessing whether it results in the maximum amount of happiness for the largest number of people. Of course, this approach is deeply problematic, too. Neither the concept of empirical observation nor the utilitarian moral maxims, including the maxim itself as well as the notoriously ill-defined notion of happiness, are empirically justifiable.[25]

Pragmatists take the failures and shortcomings of both the rationalist and the empiricist epistemologies seriously. Rather than trying to deduce, with mathematical precision and absolute validity the truth about (N 1) once and for all, they urge a more modest approach: What is true is what has passed the scrutiny of our best current methods, the dispute among the brightest minds of a generation. If these methods and disputes decide in favor of (N 1), then we should accept it until further evidence moves the weights in favor of another position. Furthermore, we should be sensitive to the many ways in which other convictions of ours influence our view on the concepts at issue.

Let us see how this could play out in some more detail for the Quinean. On a Quinean view, the truth of (N 1) cannot be decided purely *a priori* (for a pragmatist like Quine, there is no pure *a priori*). All justification, including the justification of normative claims in the legal domain, is answerable to experience, albeit perhaps in a very indirect way.

More specifically, we suggest that, on a Quinean view, normative statements such as (N 1) occupy a middle ground between (almost) purely empirical statements such as «blue there» and (almost) purely theoretical statements such as the law of the excluded middle. For instance, imagine a legal scholar who is convinced of (N 1). Then, she witnesses the devastation that (N 1) creates in a family with many children, some of which have special needs: (N 1) requires that the family be taxed at the same rate as a single person without any dependents. This experience does not logically refute (N 1). The scholar could describe her experience as mere emotional weakness which is epistemically irrelevant. In this sense, there is epistemic room for maneuver here, as is typical for statements that are at some distance from experience. Still, as the example shows, it is more realistic that experience can refute (N 1) than it is that experience refutes the law of the excluded middle.

If witnessing the plight of said family transforms the legal scholar into an ardent advocate of need-based taxation, then this is fine and well as long as she can back this up with convincing arguments in the ongoing scholarly debate. Even more, if, as a matter of fact, she does favor need-based taxation, but continues to advocate flat tax because she thinks it is better for her career and consequentially develops strong and novel arguments in favor of that position,

---

19 Ronald Dworkin, Justice for Hedgehogs, Cambridge (Massachusetts) 2011, 114.

20 Of course, this is not an empty field in legal writing as several authors have already discussed the question of whether and how values relevant for legal decisions might develop over time. See e.g. Hänni (n. 2), 202 et seq. with further references.

21 Immanuel Kant, Kritik der reinen Vernunft, edited by Jens Timmermann, Hamburg 1998.

22 John Locke, An Essay Concerning Human Understanding, edited by Peter H. Nidditch, Oxford 1975.

23 David Hume, An Enquiry concerning Human Understanding, edited by L. A. Selby-Bigge, 3rd ed. revised by Peter H. Nidditch, Oxford 1975.

24 John Stuart Mill, Utilitarianism, edited by Roger Crisp, Oxford 1998 (1861).

25 For a broader discussion of this justificatory problem for naturalists, see Reto Gubelmann, A Science-Based Critique of Epistemological Naturalism in Quine's Tradition, Palgrave MacMillan 2019.

convincing the entire field, this would also be fine from a pragmatic perspective. What matters is the free play of empirical methods and argumentative discourse – the position that comes out on top should be accepted as true, regardless of the (lack of) inner convictions of its advocates.

What is, in contrast, fundamentally threatening the progress of research is narrow-mindedness and dogmatism. The ongoing, never-ending pursuit of truth must not be blocked by anybody who claims that they have once and for all conceived the essence of tax justice correctly (the rationalist) or that they have mathematically-empirically settled that (N 1) is true as it allows for the greatest amount of happiness in the entire population (the empiricist-utilitarianist). As a consequence, the pragmatist will be a determined advocate of the free exchange of ideas, and when in doubt, she will prefer to spend her debates with colleagues that hold a view that contrasts with her own perspective.

In this manner, what is habitually referred to as one's (ultimately subjective) moral intuition might be considered a theory that can be developed and improved through experience, which includes exposing oneself to normative statements incompatible with one's own views.

In our exploratory experiment, we try to illustrate how NLP can accelerate such processes of being confronted with viewpoints that are incompatible with one's normative position with the goal of improving, perhaps even changing this normative position.

V. Neural NLP for Normative Reasoning

After arguing why, from a pragmatic perspective, it is vital for the progress and quality of research in tax law that researchers are confronted with normative positions that compete with their own views, in this section, we argue that current NLP provides the elements to build tools that could foster such confrontation.

A. General Introduction to the State of the Art in Neural NLP

There is an important distinction between symbolic AI, also called GOFAI (for «Good Old-Fashioned AI»), and connectionist AI. Symbolic and connectionist AI represent two fundamentally different approaches to design computer programs that intelligently fulfill certain tasks.[26] In brief, symbolic AI tries to address a certain task, say translating a sentence from Chinese into English, using explicit, in the typical case hand-written rules that often essentially involve some sort of logical processing of the input, for instance by using traditional predicate logic.

Connectionist models, or as they are usually called today, neural network models, proceed very much differently. Rather than explicitly specifying the rules by which a program is expected to solve a given task, neural network models consist of connected mathematical structures, so-called cells, whose parameters, also called «weights», are set during a so-called training phase. If the training phase is successful, the parameters have been set in such a way that the overall model delivers satisfactory performance at the task at hand. Roughly, this training process runs as follows:

1. Initialize all the parameters in the model with random values.
2. Let the model predict a number of values, say, the logical relationship between a number of sentence-pairs.
3. Optimize the weights of the models such as it can be expected that it will make better predictions next time.

Currently, most natural language understanding (NLU) systems are based on the so-called transformer architecture introduced by Vaswani et al.[27] NLU is the area of NLP that focuses on tasks that would, for a human being, require understanding the language in question. For instance, NLU encompasses answering questions about a given text, summarizing text, and finding logical relationships between assertions in a text. Often, models based on these architectures outperform the human baseline, as evidenced by the GLUE[28] and SuperGLUE[29] leaderboards.

Most current NLU-systems are inspired by BERT (read «Bidirectional Encoder Representations from Transformers»), an offspring of the transformer architecture.[30]

---

[26] For an introduction to GOFAI, see Margaret A. Boden, 4 GOFAI, in: Keith Frankish/William M. Ramsey (ed.), The Cambridge Handbook of Artificial Intelligence, Cambridge 2004, 89 et seq. For an introduction to connectionist AI, see Ron Sun, Connectionism and neural networks, in: Keith Frankish/William M. Ramsey (ed.), The Cambridge Handbook of Artificial Intelligence, Cambridge 2014, 108 et seq.

[27] Ashish Vaswani/Noam Shazeer/Niki Parmar/Jakob Uszkoreit/Llion Jones/Aidan N. Gomez/Lukasz Kaiser/Illia Polosukhin, Attention Is All You Need, NIPS 2017, 6000 et seq.

[28] Alex Wang/Amanpreet Singh/Julian Michael/Felix Hill/Omer Levy/Samuel Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels 2018, 353–355.

[29] Alex Wang/Yada Pruksachatkun/Nikita Nangia/Amanpreet Singh/Julian Michael/Felix Hill/Omer Levy/Samuel R. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, NeurIPS 2019.

[30] Jacob Devlin/Ming-Wei Chang/Kenton Lee/Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Lan-

BERT is being trained on general tasks, such as predicting the next word in a text, on very large amounts of data for a long time, following the general method spelled out above. Then, it is trained a second time for a much shorter time on much less data for the specific task at hand, say detecting logical relationships. The computational ratio between the former training, also called pre-training, and the latter training, also called fine-tuning, is about 3000:1. This means that the general-purpose training is clearly dominant, indicating that the models can transfer much of what they have acquired during the general-purpose training over to the specific task for which they are fine-tuned. To emphasize the importance of this first training, these models are often called pre-trained language models (PLMs).

B. An Overview of the State of The Art in Natural Language Inference

Current natural language inference (NLI) is typically conceived as a three-way classification problem. With samples such as (1), consisting of a premise (P) and a hypothesis (H), the PLMs are tasked to categorize their relationship as either one of *contradiction* (P and H cannot both be true), of *entailment* (P [logically?] implies H), or as being *neutral* (neither of the two).

(1) (P) The streets are wet. (H) It has rained.

Transformer-based PLMs fine-tuned to specific datasets perform impressively at natural language inference (NLI) tasks. The MNLI Leaderboard, for instance, shows that the top ten PLMs are without exception transformer-based. The top three are given by two different sizes of T5[31] and ALBERT[32]. Notably, in contrast to GLUE as a whole, the PLMs did not yet manage to outperform the human baseline at MNLI (as of January 14, 2022).

What is emerging as neural NLI's most pressing problem is the fact that these neural PLMs might almost outperform the crowdworker-based human baseline for the dataset on which they were fine-tuned, but perform worse than random at out-of-dataset-samples. The problem of generalization in NLI is broadly acknowledged in the literature[33]. It is generally assumed that the underlying cause of the problem of generalization is the PLMs' overfitting[34] on the training set. This overfitting, so the assumption goes, leads to the PLMs' picking up on spurious idiosyncrasies of the datasets, leading to the use of shallow heuristics and ultimately to a lack of generalization.

In summary, current NLI approaches perform impressively at samples that are very similar in structure to the ones on which they have been fine-tuned. However, their performance often drops beyond coin-toss when applied to samples affording the same logical relationships, but whose linguistic surface structure is substantially different from their training data.

For normative reasoning in legal studies, this means that it would be well possible to develop an NLP system that performs excellently at identifying logical relationships holding between different claims in a legal text—provided that there is sufficient training data (which to date is not the case).

C. Exploring PLMs' Sensitivity to Normative Claims in Legal Texts

While there is sufficient research to judge the general prospects of current NLP's ability to detect and correctly categorize logical relationships between statements in a legal text, there is no research at all to assess whether the category of normative reasoning is accessible to the models. In other words, it is not clear whether current neural NLP models would be sensitive to normative statements in a legal text. After all, there are still legal scholars that do not acknowledge the very existence of normative reasoning in their research. It is therefore a valid question

guage Understanding, NAACL-HLT 2019.

31 Colin Raffel/Noam Shazeer/Adam Roberts/Katherine Lee/Sharan Narang/Michael Matena/Yanqi Zhou/Wei Li/Peter J Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, arXiv preprint arXiv:1910.10683 2020.

32 Zhenzhong Lan/Mingda Chen/Sebastian Goodman/Kevin Gimpel/Piyush Sharma/Radu Soricut, Albert: A lite bert for self-supervised learning of language representations, International Conference on Learning Representations 2019.

33 See Xiang Zhou/Mohit Bansal, Towards robustifying nli models against lexical dataset biases, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020, 8759–8771; Ronan Le Bras/Swabha Swayamdipta/Chandra Bhagavatula/Rowan Zellers/Matthew E Peters/Ashish Sabharwal/Yejin Choi, Adversarial filters of dataset biases, arXiv preprint arXiv:2002.04108 2020; Prasetya Ajie Utama/Nafise Sadat Moosavi/Iryna Gurevych, Towards debiasing nlu models from unknown biases, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020, 7597–7610; Dimion Asael/Zachary Ziegler/Yonatan Belinkov, A generative approach for mitigating structural biases in natural language inference, arXiv preprint arXiv:2108.14006 2021; Rabeeh Karimi Mahabadi/Yonatan Belinkov/James Henderson, End-to-end bias mitigation by modelling biases in corpora, arXiv preprint arXiv:1909.06321 2019; and Jean-Philippe Bernardy/Stergios Chatzikyriakidis, What kind of natural language inference are nlp systems learning: Is this enough?, ICAART (2) 2019, 919–931.

34 See Ian Goodfellow/Yoshua Bengio/Aaron Courville, Deep learning, Cambridge (Massachusetts) 2016.

whether pre-trained language models (PLMs) are able to identify such statements.[35]

To explore the promises of NLP in navigating normative judgments in tax law, we have conducted an exploratory clustering experiment. In a clustering task, the goal is to group together samples into the same cluster that belong together. In our experiment, the goal was to group together these normative statements that express the same basic normative outlook.

For the purposes of this exploratory experiment, we have identified four normative stances, namely libertarianism, Rawlsianism, as well as procedural and deontological views. The following description of these categories remains sketchy, as its sole purpose here is to give an idea of the task presented to the clustering algorithms tested.[36]

For libertarianism, the essential idea is that the market outcome regarding income and wealth distribution is just and deserved and, therefore, taxation should not lead to redistribution. As famously held by Nozick: «*taxation of earnings is on a par with forced labor*»[37].

Rawlsian arguments, in contrast, are arguments in favor of a justified level of inequalities. This includes arguments in favor of redistribution, e.g., through progressive taxation. Rawlsian arguments focus on the redistributive effect of the tax system in the sense that the tax system may help in reducing unjustified inequalities to a level which is in favor of the least well-off in a society.[38]

The term deontological ethics covers a broad variety of approaches.[39] For the purpose of the present paper, we consider arguments as deontological arguments if they focus on the treatment of the taxpayer as opposed to any effects of this treatment, say the (re)distribution of the income within a society. The category helps us to cover the widespread argument that taxpayers should be treated equally (i.e., horizontal equity).

Procedural arguments, finally, highlight that just tax laws are the outcome of free deliberative debate about the main design elements of the societal structure. This includes, for instance, a Habermasian approach[40] aimed at achieving a just societal structure based on a democratic decision-making process (with regard to this emphasis on procedures, procedural positions therefore resemble pragmatism in epistemology). Hence, for the purposes of our exploratory experiment, procedural arguments are understood as arguments either in favor of or against a certain tax policy suggestion based on an exchange of ideas in a deliberative process.

For the clustering, we manually grouped 10 samples of each of the four normative categories identified. This yields a total of 40 samples that were submitted to the clustering experiment. The samples are all grammatical sentences taken from journal articles, and their categorization was conducted by an expert in the field. Then, a trained philosopher without specific expertise in the field went through the examples and annotated any disagreements. Where the disagreements could not be resolved by discussion, a different sample, whose categorization was uncontroversial, was chosen.

We explored the promises of using embeddings produced by pre-trained transformer-based PLMs that are all cognates of the very influential BERT architecture[41] together with two different clustering algorithms with various configurations. In these experiments, the PLMs delivered so-called embeddings, which were then grouped by the different kinds of clustering algorithms. Roughly, the embeddings are high-dimensional vector representations of the normative claims. Such representations have been shown to represent semantic as well as syntactic features of the words or sentences that they represent.

In our clustering experiment, these embeddings are of central importance. If embeddings representing sentences categorized by the experts as Rawlsian, deontological, etc. are close together in the vector space, this will greatly facilitate adequate clustering.

We varied a number of variables that are available when clustering sentences, including the embeddings used, the clustering algorithm applied as well as the parameters that these algorithms take. Overall, this has resulted in 875 different configurations, which were all evaluated using our test set.

In brief, our results show that the best-performing systems manage to deliver impressive results. Figure 1 shows

---

35 For the technical details of our experiment, see Reto Gubelmann/Peter Hongler/Siegfried Handschuh, Exploring the Promises of Transformer-Based LMs for the Representation of Normative Claims in the Legal Domain, arXiv:2108.11215 2021.

36 For details on the normative categories as well as on the dataset built on it, see Reto Gubelmann/Peter Hongler, Natural Language Processing – A Game Changer for Normative Reasoning in the Legal Discipline?, IFF Working Paper 2021.

37 Robert Nozick, Anarchy, State, and Utopia, New York 1974, 169.

38 John Rawls, A Theory of Justice, Cambridge (Massachusetts) 2007.

39 See e.g. Larry Alexander/Michael Moore, Deontological Ethics, The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), edited by Edward N. Zalta, Internet: https://plato.stanford.edu/archives/sum2021/entries/ethics-deontological/ (last consulted on 18.8.2021).

40 Jürgen Habermas, Theorie des kommunikativen Handelns, Berlin 1995.

41 See above, section V.B.

the performance of a clusterer that is based on sbert-embeddings[42] and that uses a k-means clustering algorithm[43] with the number of clusters set to 4 (for k-means, it is necessary to specify in advance how many clusters the clusterer should identify).

Figure 1 shows clearly that the system has largely clustered correctly: in each of the four groups identified, one of the four categories dominates clearly. For instance, the cluster in the second row has 9 total members, 8 of which belong to the Rawlsian category.

largest membergroup	total members	homog.	wh	Img label
8	9	0.89	0.20	Rawlsian
9	13	0.69	0.23	Libertarianism
8	9	0.89	0.20	Procedural
6	9	0.67	0.15	Deontological

Figure 1: The clustering results of the best-performing configuration in our experiment.

This result is very encouraging. If we consider that the models used in these experiments were not fine-tuned to the specific genre, namely legal texts, let alone to the task at hand, namely clustering normative statements, then it becomes clear that this marks merely the lower boundary of what is possible with more substantially adapted and optimized models.

More broadly speaking, it shows that state-of-the-art PLMs are sensitive to normative categories: As figure 1 shows, the best-performing model (which is notably a multilingual model, opening interesting venues for transferring the representative capacities of the model to German) correctly identifies the four normative categories set-up by the experiment, namely Rawlsian, Libertarianism, Procedural, and Deontological. This was not clear from the start: It would have been just as possible, for instance, that the models lump together Rawlsian and Deontological statements (as Rawlsianism can be conceived as a kind of deontological position), splitting up the rather diverse procedural class—or that they are simply insensitive to the normative categories, instead using surface phenomena such as sentence length for clustering. As a matter of fact, however, the best-performing systems were able to identify these four classes. Furthermore, the classes are surprisingly homogeneous with regard to the normative viewpoints of their members. We take this as strong evidence that current neural NLP approaches are sensitive to normative statements in legal research texts.

Together with the state of the art in natural language inference sketched above[44], this suggests that current NLP provides the elements to build tools that are able to improve normative reasoning and with it the overall quality of legal research in the manner specified above from a pragmatic background[45], namely by contradicting researchers illuminatingly on normative territory.

VI. How NLP Can Improve Normative Reasoning in Legal Studies: Getting Specific

As we have stated above, we claim that NLP has the potential to qualitatively improve research in legal studies. While it has become almost commonplace that it will lead to quantitative gains, for instance by reviewing a huge amount of case law in order to figure out whether there has been any similarity to the disputed case,[46] the idea that it could also qualitatively improve research is new to many.

To support our claim, we have chosen the area of normative reasoning, arguably one of the least automation-friendly areas in legal studies. We have defended our case on philosophical-conceptual grounds (defending a pragmatist epistemology of normative reasoning above, section IV.) as well as on technical grounds (section V.), introducing the state of the art in NLI as well as experimentally verifying the sensibility of state of the art systems for normative categories in legal reasoning.

The core benefit that we think can be expected from NLP is, briefly, that it is able to contradict us in insightful ways. If we accept the basic pragmatic conviction that insights in normative matters are not obtained by individual scholars that deduce the truths either from a priori principles or from observations, but rather by the methodical, rule-governed process of debate, then we should strive to consider as many well-thought-through positions and arguments as possible. We should accept as justified these positions that have stood their ground in these debates.

More specifically, we think that NLP can both help to make normative positions and arguments visible as such,

---

42 See Nils Reimers/Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 2019, 3982–3992.

43 Stuart Lloyd, Least squares quantization in PCM, IEEE transactions on information theory, vol. 28, issue 2 1982, 129–137.

44 See section V.B.

45 See section IV.

46 See e.g. Internet: https://hai.stanford.edu/news/natural-language-processing-ready-take-legal-hearings (last consulted on 18.8.2021).

and that it can also efficiently present us with high-quality arguments that contradict our own considered positions so as to allow us to learn from these opposing viewpoints and hence improve our own stance. These two benefits can have a number of very specific effects that will result in a qualitative improvement of normative reasoning in the legal domain. In particular, we see promise of progress in the following four areas.

First, it is often argued that impartiality (in this context also called objectivity) as part of a legal interpretation process is difficult to achieve if not illusionary, as each of us has a different preunderstanding, i.e., an understanding of a term out of the context of a specific legal provision. In principle, such criticism is justified. However, we suggest that the more we are faced with statements and arguments which might deal with a similar interpretation problem from a different normative point of view, the more we will be inclined to take these points of view into account and hence produce a more impartial and in this sense more objective assessment. We have already sketched how this ties in with a Quinean and generally pragmatic epistemology of normative reasoning.

Second, it has become common to have several commentaries about the same law. However, some commentaries remain rather superficial; normative considerations are hardly ever included, even if they would be central to the topic. One reason is that it is extremely time-consuming to review all existing opinions about a certain topic. Therefore, NLP might allow authors to develop a more in-depth line of argumentation in the same amount of time as NLP might allow a pre-selection of reading material.

Third, those who tend to refuse any debate about essentially contested concepts such as fairness or justice might find themselves in a worse position, as NLP is able to detect and categorize normative statements within seconds. Therefore, NLP will help to detect normative statements in a legal text, which will compel authors, especially those who tend to deny any value-based reasoning in their texts, to deal with values, as a computer system might be able to calculate to what extent an argument contains a value-based judgment.

Fourth, and this is likely the most ambitious expectation: NLP might even allow to assess scholarly writing in an objective manner in the sense of measuring the depth of arguments and thereby provide lawyers with a tool to assess the depth of an entire article.[47]

VII. Conclusion

In this article, we have argued in a loosely Quinean way that research in tax law, and in the theories of tax justice underlying this research, would greatly benefit from engagement with the various normative points of view that are in play in these debates. However, to engage them, they must be made explicit. We have suggested that there is great potential in NLP to achieve this, that is, to make these normative backgrounds explicit that are presupposed by tax legal arguments, and to improve legal research by confronting scholars with normative points of view that are fundamentally opposed to their own.

[47] In certain contexts, and with slightly different purposes, such systems are already available, compare Thiemo Wambsganss/Christina Niklaus/Matthias Cetto/Matthias Söllner/Siegfried Handschuh/Jan Marco Leimeister, AL: An Adaptive Learning Support System for Argumentation Skills, ACM CHI Conference on Human Factors in Computing Systems 2020.