On What it Means to Pay Your Fair Share: Towards Automatically Mapping Different Conceptions of Tax Justice in Legal Research Literature

Reto Gubelmann and **Peter Hongler** and **Elina Margadant** and **Siegfried Handschuh**
University of St.Gallen
Dufourstrasse 50
9000 St.Gallen
{reto.gubelmann,peter.hongler,elina.margadant,
siegfried.handschuh}@unisg.ch

Abstract

In this article, we explore the potential and challenges of applying transformer-based pretrained language models (PLMs) and statistical methods to a particularly challenging, yet highly important and largely uncharted domain: normative discussions in tax law research. On our conviction, the role of NLP in this essentially contested territory is to make explicit implicit normative assumptions, and to foster debates across ideological divides. To this goal, we propose the first steps towards a method that automatically labels normative statements in tax law research, and that suggests the normative background of these statements. Our results are encouraging, but it is clear that there is still room for improvement.

1 Introduction

Disagreements about normative claims are notoriously hard to resolve, and in some cases, they are even hard to recognize as such. For instance, consider (1). Do you think that a tax system that follows this principle is just?

(1) To be just, a taxation system must tax people with the same income equally.

Example (1) illustrates what we mean by a normative claim: A moral judgment of some kind, that is, an assertion that something is either morally right or wrong. As we restrict our scope to tax law, the normative claims that we are interested in pertain to moral judgments of specific tax systems. Hence, while example (1) counts as a normative claim, example (2) does not count. While the latter is also about tax law, it does not make a claim about what is just or unjust in this domain, but rather what is legal.

(2) It is illegal not to pay one's taxes.

In the discussion on tax justice, claims of the kind of (1) are regularly made, and even more often they figure implicitly in the arguments of legal scholars. For example, consider (3), which does not make an explicit claim about what is just in taxation matters, but which implicitly presupposes an idea of the kind expressed in example (1). In the worst case, adherents of different normative positions will retreat into their normative bubbles and hence permanently hinder any truly rational debate about these topics.

(3) Taxation cannot consider the needs of the individuals or their dependents, as this would lead to people with the same income being taxed at a different rate.

To move towards improving this situation, we explore the use of state-of-the-art PLMs to detect and to classify normative statements in tax law research texts. More specifically, using a variety of classifier configurations, we first explore the sensibility of state-of-the-art PLMs for different normative backgrounds in a hyperparameter search experiment. Second, we use the configurations that have shown to perform best to iteratively develop a dataset as well as a method that both identifies normative statements and that classifies them into five distinct normative categories. Finally, we validate our results with the help of two experts without any previous knowledge of the project.

We make two contributions to the field. First, studying a domain within legal NLP that has so far remained entirely uncharted, we provide specific recommendations and insights for further research in this area. Second, we publish a high-quality, expert-verified dataset for this domain that is of considerable size given the complexity of the task.

We note that our task and domain differ sub-

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 12 - 30
December 8, 2022 ©2022 Association for Computational Linguistics

stantially from, and hence nicely complement first-order legal NLP tasks: rather than analyzing specific rulings, provisions, or contracts, our target is the second-order discussion about which kinds of provisions, rulings, and contracts might be just, which means that the content, vocabulary and goals of our target texts will differ accordingly. Our domain also complements studies of the normative attitudes used to describe individual moral stances, such as in moral foundations theory, or in human values approaches, see Kiesel et al. (2022) and Hoover et al. (2020): Our research focuses on a discussion that belongs to political philosophy, centering around the question what constitutes a just system of taxation, rather than analyzing the moral motivations of individuals to adopt one position rather than another.

While these second-order discussions about tax justice are clearly separable from the first order ones, the former directly influence the latter. If a judge subscribes to the libertarian view that income taxation is "on a par with forced labor" (Nozick, 1974, 169), then her rulings will show much more sympathy for individuals who try to avoid paying taxes by all means. In contrast, if she subscribes to a more Rawlsian view that mandates redistribution of wealth insofar as it constitutes unjustifiable inequalities, she will have much less sympathy with wealthy individuals who are optimizing their tax bills.

The task in focus of this article is both challenging and important. It is challenging because recognizing the specific normative background of a statement such as (1) and (3) requires expert knowledge, and even with such expert knowledge, genuine uncertainties remain in some cases. More fundamentally, it means that the very definition of the categories as well as the identification of the first samples that fall under these categories requires expert knowledge from legal studies. As a consequence, the present project is interdisciplinary throughout: only a combination of expertise in legal studies and NLP can achieve progress on this topic. In our second experiment, we address this challenge by iteratively combining expert input and classifiers in a bootstrapping procedure.

Furthermore, considered from a technical perspective, the amount of lexical overlap is substantially higher than in typical clustering or classifying settings, say, in classical word-sense-disambiguation (WSD) tasks (see Navigli 2009 for a survey), where the method has to distinguish entirely different senses of words such as "bank". This is because different normative conceptions of tax justice do not constitute fully-fledged cases of ambiguities: if adherents of two different normative theories debate tax justice, they might disagree strongly on the correct conception of tax justice. However, unlike in bank-cases of ambiguity, both mean to capture the same idea.

In the pertinent philosophical and linguistic literature, such concepts are called "essentially contested concepts". The conception was first proposed by Gallie (1955), for recent discussions see Collier et al. (2006) and Rodriguez (2015). According to this conception, concepts such as TAX JUSTICE are such that essential parts of their meaning are disputed. And the reason for the dispute is that the disagreement is due to larger-scale differences in worldview.

The task is important because the subject matter that is addressed in such normative arguments is of central importance for liberal democratic societies. The politically crucial questions of liberal states are of essentially contested nature. What counts as a just taxation system directly influences the lives of the members of that society. Hence, providing support to navigate such normative landscapes is of central importance for liberal democratic societies.

2 Related Research

We focus on two areas of related research: the work on word- and sentence-embeddings that we use to represent statements, and legal NLP, the subdomain of NLP that is concerned with legal texts.

We emphasize two aspects that separate our focus from that of current related research. First, the vast majority of research in legal NLP focuses on first-order legal texts, that is, specific provisions, court decisions, or contracts. In contrast, we focus on second-order legal texts, that is, on research literature about such provisions or court decisions. Second, to date, there simply is no research that focuses on normative positions within the legal domain. These two distinguishing characteristics force us often to resort to generic approaches.

We use three different kinds of embeddings for our experiments; two of them are based on the transformer architecture (Vaswani et al., 2017), the third kind consists of classic distributional embeddings. First, we use embeddings generated by **word-based PLMs**, namely bert-base-cased and bert-large-cased (Devlin et al., 2019) as well as roberta-large (Liu et al., 2019). Among this category, we include legal-bert, a model specifically designed for first-order legal texts (Chalkidis et al., 2020).[1]

Second, we test a number of **sentence-based PLMs**, namely SBERT-Models (Reimers and Gurevych, 2019), as initial explorations showed that they perform clearly best. These SBERT-Models are based on a variety of transformer-based PLMs (in addition to the classical BERT and RoBERTa, these are mpnet, Song et al. 2020, distilroberta, Sanh et al. 2019, AlBERT, Lan et al. 2019, and minilm, Wang et al. 2020). SBERT-Models are optimized for sentence-level comparison of embeddings via geometric similarity or distance measures such as cosine similarity.

Third, we use non-transformer-based, distributional **classical word embeddings**, namely GloVE (Pennington et al., 2014) and Komninos (Komninos and Manandhar, 2016) for purposes of comparison.[2]

For classification, we use support-vector-machines ("SVMs", Boser et al. 1992); SVMs systematically try to find the optimal hyperplane separating samples of different categories. We use the scikit-learn implementations of all the clustering and classification algorithms used in this study, see Pedregosa et al. (2011).

Dale (2019) shows that NLP has been used in the legal domain since the 1960ies, with the size and the financial significance of the legal business seemingly creating a perfect environment for the development and application of domain-specific NLP methods. However, as Tang and Clematide (2021) detail, the legal domain poses specific challenges, among them the unusual length of typical legal documents, a jargon that differs on the lexical and syntactic level from standard English, domain-specific notions of relevance, and the high cost of obtaining high-quality labelled data (legal experts are expensive).

These challenges explain why many core tasks in legal NLP are still unsolved. Perhaps most prominent among them is the task of finding relevant legal documents (i.e., codified legal texts as well as authoritative court decisions) given a specific query. Thus, Chalkidis et al. (2020) systematically investigate good practices for training transformer-based PLMs that perform well in typical first-order legal tasks (classification of laws and court decisions as well as named-entity recognition in contracts). Soh et al. (2019) evaluate different methods to classify Singapore supreme court decisions according to the legal area involved, finding that rather simple combinations of latent semantic analysis and support vector machine to perform equally well as state-of-the-art PLMs. With their survey, Chalkidis and Kampas (2019) provide embeddings based on the word2vec method (Mikolov et al., 2013) that are derived specifically from court decisions and legal provisions.

As the specific idiom of legal texts is challenging already within English, multilingual research is all the more challenging. There has been some research with regard to German. Wrzalik and Krechel (2021) present a German dataset for information retrieval, Niklaus et al. (2021) focus on judgment prediction of the Swiss federal court, whose rulings are translated in German, French, and Italian, being all official languages in Switzerland.

Regarding the specific area of tax law, Ash et al. (2021) present a novel approach to identify legal documents as belonging to the field of tax law, and within this field, classifying them into specific subclasses, such as personal income or sale. As a consequence, the structure of their classifying task is somewhat similar to ours: We, too, are interested in first identifying normative statements as such and then assigning them to specific normative positions. Note, again, however, that this study also focuses on first-order tax law provisions rather than on legal research articles reflecting such tax law provisions, which is our focus.

Our research shows some connections with the ongoing discussion about so-called open-textured concepts. According to Rissland and Skalak (1989, 525), open-textured concepts are such that they cannot be defined by necessary and sufficient conditions. This category is obviously broader than the one of normative concepts and statements in focus here: Rissland and Skalak (1989, 525) mention "meeting or dealing" and "contract" as examples.[3]. For an early attempt to tackle reasoning with such

[1] These models were downloaded from huggingface.co, see Wolf et al. (2019).

[2] The SBERT- as well as the classical models were obtained from https://www.SBERT.net/docs/pretrained_models.html.

[3] Indeed, building on Ludwig Wittgenstein's conception of family resemblances ("Familienähnlichkeiten"), one could argue that all concepts with the exception of very few, highly artifical cases are open-textured, as it is usually not possible to give a definition whose parts are individually necessary and jointly sufficient for concept application in all relevant contexts. See Wittgenstein (2006/1953).

concepts, see Sanders (1991). A recent categorization of regulation with an eye towards their potential to be processed automatically point out that such open-textured concepts are a considerable obstacle to such automatic processing (Guitton et al., 2022).

The essentially contested concepts that are often at the core of normative claims in focus of this paper can be seen as a specific species of open-textured concepts, namely those that resist any simple resolution of their open-texturedness due to being conceived very diffently from very different comprehensive worldviews.

3 Datasets

As we are interested in normative positions within research discussions in tax law, all of our datasets consist of statements from such research articles. The full references of these articles are listed in the appendix, section A. In addition to these research texts, we had to develop suitable classes to categorize the normative statements. Our tax justice expert supervised the development of five normative positions that are particularly prominent in the field. These five positions constitute the categories for our experiments.[4] In the following, we first introduce these five normative categories. Then, we detail specifics of the datasets used for each of our three experiments.

According to the so-called *Deontological View*, a tax policy proposal is just if it focuses on the treatment of the taxpayer and not on the distribution of the income within a society. Hence, according to the Deontological View, a tax provision is just if it conforms to basic moral principles, such as the fundamental equality of all human beings. In this sense, example (1) expresses a Deontological View.

According to the *Rawlsian View*, a tax system is just if it would be chosen by individuals that are under Rawls' famous veil of ignorance. Under this veil, individuals do not know their educational, financial, social, or any other position in the society whose tax system they are supposed to judge. It is generally agreed that such individuals would favor tax systems focused on equality and on the eradication of unjustified inequalities.

A tax provision is just if it results from good, democratically grounded processes – this is the gist of the *Procedural View*. Such view includes positions that argue for a certain tax policy proposal based on a discussion or debate about the arguments against and in favor of such a proposal.

The fourth theory used in this article is the *Libertarian View*. According to it, taxation should be kept at a minimum in general, as it is considered illegitimate in all but a few cases, mostly where it is necessary to allow a minimal state to function. Libertarians tend to view market outcomes as just and therefore any kind of redistribution as unjust.

The fifth and final normative viewpoint to be included in this study is *Utilitarianism*. According to it, we should develop a taxation system that results in the maximal increase in the overall population's happiness, or welfare. This means that, according to Utilitarianists, it is permissible that individuals are treated unequally if this implies a net benefit in welfare or happiness for the entire population.

Table 1 shows the names of each of the categories, including the None-Class with a typical example.

Specifics for Experiment 1 For the hyperparameter search, we asked the expert to manually find 35 samples of each of the five normative categories identified in publications in peer-reviewed journals from the legal domain, yielding an evenly distributed dataset of 175 samples. As we expected that most of the sentences that the classifier would encounter are not expressing a normative perspective, we then added 1708 non-normative statements in the following way. Using an sbert-sentence-embedding-model, we computed the centroid of all sentence embeddings of these 175 statements. Then, we ran this over all sentences from the corpus of bootstrapping loop 1, yielding a list of sentences with the cosine between their embeddings and our centroid. From this, we selected 523 statements with a cosine below 0.2, 310 with a cosine between 0.2 and 0.6, and then 875 with a cosine higher than 0.6. An expert in the field checked all 1708 statements to ensure that they are indeed not normative in our sense. The choice of distribution of our nonnormative samples is based on the hypothesis that the most difficult decisions to make for the classifiers are those where the overall similarity of the embedding to the centroid is high, while the statement is clearly not normative.

[4] Note that we did not find a single instance where one sentence explicitly expressed views that belong to two different categories. What we did find, of course, are cases where it is not clear to which category it belongs.

Category	Example
Deontological	Max burdens should bear similarly upon persons whom we regard as in substantially similar circumstances, and differently where circumstances differ.
Libertarian	the anti-progressive tax argument is often characterized as an argument that every person has a responsibility to take care of himself, and no one, including the wealthy, has an obligation to assist those in need.
Procedural	For Locke himself, the key institutional requirement was that taxes should not be levied except by "the consent of the people," which he understood as "the consent of the majority, giving it either by themselves, or their representatives chosen by them."
Rawlsian	The increasing inequality of market income can be significantly ameliorated by the redistributive effect of the tax transfer system, if it is appropriately targeted.
Utilitarian	Efficiency analysis looks to overall social welfare as a measure of a tax's virtue.
None	An income tax can be used to redistribute taxable income.

Table 1: The five normative categories used in the experiments including the None-Class with typical examples.

Loop	Single-gate	Dual-gate
0	175/1708	175/1708
1	310/1767	292/2091
2	435/1792	452/2172
3	686/1892	709/2415
Combined Final DS	937/2194	

Table 2: Listed in loops 1-3 are the resulting, expert-reviewed datasets after each loop (Normative/Nonnormative samples). Dataset at loop 0 represents the input to bootstrapping loop 1 that is equivalent to the dataset used in experiment 1. For the meaning of "single-gate" and "dual-gate" see below, section 4.2.

Specifics for Experiment 2 In our iterative bootstrapping experiment, we used separate texts as sources for the initial expert-compiled dataset as well as for each of the three bootstrapping loops (for references, see the appendix, section A). Note also that the training datasets for the classifiers grow with each further bootstrapping loop taken, as we include the corrected output from the previous bootstrapping loop in the training dataset for the next one. Table 2 gives the details of the datasets, as they evolved through the bootstrapping process.

Specifics for Experiment 3 We presented our external expert annotators with a dataset of 650 samples in total. This consists of evenly distributed samples (i.e., 130 samples of each of the five categories) from the final dataset resulting from experiment 2. That is, it contains samples of three different origins: (1) samples that are directly extracted from the texts by a human, (2,3) samples that have been suggested by one of our two classifying methods and then reviewed by a human expert.

We publish the final dataset, as well as other material that might be useful to the community, on GitHub.[5]

4 Experiments

The goal of our experiments is twofold (see above, section 1). First, using a human-in-the-loop method, we want to develop a high-quality dataset of normative statements from tax law that can serve as the basis for further studies of this and related fields by the community. Second, we want to assess whether current models, both generic ones and others fine-tuned to the legal domain, are able to map the subtle differences that exist between these different normative perspectives on tax law. Given that the field that we are working in is entirely uncharted, we believe that this double aim maximizes the benefit to the research community, and we have designed the experiments accordingly.

4.1 Experiment 1: Hyperparameter Search

The goal of this first experiment consists in finding the best hyperparameters for our main experiment 2. We tested a number of support vector machines, varying the usual hyperparameters and combining this with a total of 23 different pre-trained language models (PLMs). We tested three different kinds of PLMs. First, different transformer-based word-based models, including generic pre-trained BERT and RoBERTa as well as a model specifically developed for first-order legal texts, legal-bert. Second, we tested a number of transformer-based sentence-bert models, and third, we included two pre-transformer distributional models. For refer-

[5] Please consult this repository.

ences, see above, section 2, for details of the models as well as the configurations tested, see the appendix, section B.

Furthermore, we tested the configurations on two different tasks. In the first task, the classifiers had to categorize a dataset of 175 samples, evenly distributed across the five categories, into one of the five categories (called the “5cat task”). In the second task, the classifiers had to categorize a dataset of 1883 samples into normative and nonnormative, with 175 (the same that were used for the first task) being normative, and 1708 being nonnormative (called the “Norm task”). This uneven distribution is intended to model the actual task in the wild, where we expect the clear majority of sentences encountered by the classifiers to be nonnormative on our reading.

Overall, we tested 1380 different SVM-configurations per task, saving the best performing SVM-hyperparameter-setup per model.

4.2 Experiment 2: Bootstrapping a Classifier and a Dataset

In this second experiment, we employed the two best-performing PLM and SVM configurations from experiment 1 to iteratively develop a classifier as well as a dataset. For details of the configurations, see the appendix, section C.

We start out with the dataset used from experiment 1, that is, with 175 normative sentences that are evenly distributed among the five classes as well as 1708 nonnormative sentences. This dataset is then used to train a classifier, which is run on a set of texts, resulting in predictions, which are then reviewed by an expert. These predictions, with their labels corrected by the expert, are then merged with the training dataset from this bootstrapping loop and together serve as the training dataset for the next bootstrapping loop, etc. Overall, three bootstrapping loops were executed.

We conducted these three bootstrapping loops with two different SVM methods, calling them single-gate and dual-gate. The first, called single-gate, is a straightforward classifier conceiving nonnormative sentences as a sixth category to be classified by the classifier. Here, we were using a one vs. one scheme, meaning that we are in fact training $\frac{Nx(N-1)}{2}$ classifiers, resulting in 15 classifiers. The classifier then predicts the one class that wins the most 1:1-duels. However, we hypothesized that this procedure would be not only computationally expensive, given the large size of one of the classes, namely the None-class, but also yielding bad predictions, as the None-class is nearly 40 times larger than the other classes.

We therefore also used a method that we call dual-gate method. Here, a first SVM decides on whether the sentence under consideration is normative in our sense or not (here, the normative training split is less than 10 times smaller than the None class). Then, a second gate (hence the name), consisting of 10 1:1-SVMs, classifies sentences that are normative according to the first SVM into one of the five normative classes. In this way, we employ a one vs. rest approach to distinguish normative from nonnormative sentences and a one vs. one approach to classify normative ones into their separate categories. This way, we hoped to maximize accuracy and beat the standard single-gate 1:1-approach.

4.3 Experiment 3: Annotation by Two Uninvolved Experts

In this experiment, we get an external and intersubjective view on the results of experiment 2 by having two external annotators review the dataset described above (section 3). Two ideas were guiding our design of this experiment. First, we wanted to make sure that the results of experiment 2 are not overly optimistic because our expert annotator is biased towards, as it were, annotating such that our experiments become a success. We cannot rule this out with an annotator as ours that is quite involved in our experiments. Therefore, we chose two annotators that have no involvement whatsoever in the study.

The second motivation of this third experiment was to obtain a reliable figure on the intersubjectivity of the annotations that our internal expert annotator produced. A high inter-annotator agreement would mean that many of the samples can be rather clearly assigned to a category, despite the intricacies of our subject matter.

As a consequence, we recruited two external annotators, both advanced undergraduate or graduate students in philosophy, without any previous knowledge of our project. We give the precise instructions given to the annotators in the appendix, section D. The annotators were given the opportunity to annotate “OTHER” when they were fully certain that the sample at issue, while being normative, did not fit any of the categories in focus.

Modelname	Type	5cat
pp-ml-mpnet-base-v2	sbert	87%
pp-mpnet-base-v2	sbert	85%
nli-mpnet-base-v2	sbert	84%
stsb-roberta-base-v2	sbert	83%
stsb-droberta-base-v2	sbert	83%

Table 3: Models and modeltypes used for the five best performing classifiers in the 5cat task. "pp" = paraphrase, "ml" = multilingual, "droberta" = distilroberta, "du" = distiluse, "awe" = average_word_embedding.

Furthermore, the annotators were not given any information on the three different subsets involved in the experiment, nor were they shown the predictions issued by the methods, or the categorization by our internal expert annotator – all with the goal of removing any possible bias that the annotators could develop.

5 Results

5.1 Experiment 1

The results of the two different classification tasks can be seen in tables 3 and 4 with "5cat" referring to the task of classifying samples into the five normative categories (most frequent sense baseline 20%, table 3) and "Norm" referring to that of distinguishing between normative and non-normative samples (most frequent sense baseline 91%, table 4; all results from all models are listed in the appendix, table 6). What is evident in the former case is that the models all perform rather well. Even the model that performed worst, `legal-bert-base` reached 74% accuracy. The best performing classifier is based on sentence-bert embeddings, and it is a rather small multilingual model: `paraphrase-multilingual-mpnet-base-v2`. The first classifier using classical word-embeddings employs `roberta-large`, and it loses no less than 5% to the best classifier.

The results of the Norm task differ in several aspects (see table 4). First, we find that the best classifier is indeed based on classic word-based embeddings delivered by `roberta-large`. It beats the first sentence-bert-based classifier by 3 percentage points. Given that the most frequent sense baseline is at 91%, these three percentage points are a considerable difference. Furthermore, overall, only 6 of 23 embeddings manage to ground

Modelname	Type	Norm
roberta-large	avword	98%
pp-droberta-base-v2	sbert	95%
nli-droberta-base-v2	sbert	95%
stsb-roberta-base-v2	sbert	94%
stsb-droberta-base-v2	sbert	94%

Table 4: Models and modeltypes used for the five best performing classifiers in the Norm task.

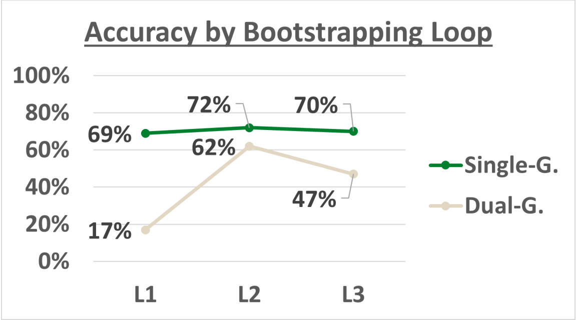


Figure 1: Overview on the performance of the two methods through the progress of experiment 2, Li referring to loop number i.

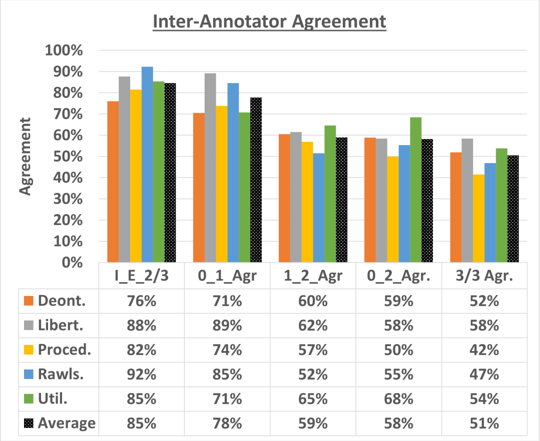
classifiers that beat the baseline, whereas in task one, all of them achieved this by a margin of 54 percentage points.

As a consequence, we decided to run the bootstrapping loops with the two different methods described above, section 4.2. We chose this strategy because we were impressed at the challenge that the task of distinguishing normative from nonnormative sentences posed to the classifiers, and we thought it necessary to have an SVM that can harness the full information contained in the samples of all normative categories to mark a good geometrical divide between these samples and the nonnormative ones.

5.2 Experiment 2: Bootstrapping a Classifier and a Dataset

An overview on the results of the three bootstrapping loops can be found on figure 1. Overall, it shows that the single-gate method outperforms the dual-gate method, despite our worries due to the large imbalance of the dataset. In terms of accuracy, it beats the dual-gate method throughout.

Table 2 (see above, section 3) shows the evolution of the two datasets through the bootstrapping process. It shows a steady growth of both normative samples belonging to one of the five categories as well as nonnormative samples through the loops,



	I_E_2/3	0_1_Agr	1_2_Agr	0_2_Agr.	3/3 Agr.
Deont.	76%	71%	60%	59%	52%
Libert.	88%	89%	62%	58%	58%
Proced.	82%	74%	57%	50%	42%
Rawls.	92%	85%	52%	55%	47%
Util.	85%	71%	65%	68%	54%
Average	85%	78%	59%	58%	51%

Figure 2: Results of experiment 3, annotator 0 is our internal expert, 1 and 2 have been recruited externally. “I_E_2/3” is the percentage of samples where our internal annotator agreed with at least one external annotator.

with the dual-gate method resulting in a slightly larger dataset with regarding to normative samples and a much larger one with regard to nonnormative ones. Furthermore, the fact that the dataset from dual-gate SVM after loop 3 is 76% the size of the final combined dataset shows that the overlap between the true positives from the two methods is quite large.

5.3 Experiment 3: Annotation by Two Uninvolved Experts

The results from our third experiment are displayed in figure 2 (we also give Cohen’s Kappa as well as inter-annotator variation by source in the appendix, section D). It shows that, in total, 85% of all of the classifications are supported by a 2/3-majority-vote, with one of the voters being external, one internal (to avoid falsely capitalizing on the two external annotators agreeing on a different label than our internal annotator, we focused on this restricted 2/3-agreement figure, abbreviated by “I_E_2/3”). This means that two out of three annotators independently identified the same category out of a choice of five categories. Annotator 0 is our internal annotator, annotators 1 and 2 are external ones. Figure 2 shows, for instance, that annotator 2 disagrees relatively often with annotators 0 and 1: while 0 and 1 agree in 78% of cases, this figure drops to about 60% if annotator 2 is involved.

6 Discussion

6.1 Experiment 1: Hyperparameter Search

The results of the hyperparameter search experiment are encouraging. For both tasks, our search has identified very promising candidate combinations of embeddings and SVM-configurations. It might be surprising that a multilingual and rather small model – mpnet-base is of the same category as bert-base, having 110M parameters – outperforms the large and monolingual models. This, however, dovetails nicely with the rankings on the SBERT-page for clustering.[6] We hypothesize that, for our task, the larger models overfitted to nonnormative settings, and hence generalized worse to this novel task.

This finding that larger models perform worse at a natural language understanding task is not entirely without precedent. For instance, researchers at DeepMind find that larger models do not necessarily perform better at natural language inference. The large study by Rae et al. (2021, 23) strongly suggests that, in the words of the authors, “the benefits of scale are nonuniform”, and that logical and mathematical reasoning does not improve when scaling up to the gigantic size of Gopher, a model having 280B parameters.

6.2 Experiment 2: Classifying

We make three observations on the results of experiment 2. First, the single-gate method outperforms the dual-gate method in terms of accuracy, but the difference decreases after bootstrapping loop 1 (see figure 1). In this loop 1, the accuracy of the dual-gate method is at 17%, whereas the single-gate method reaches 69%. This also means, given our set-up, that the dual-gate method receives a lot of high-quality false positives to use in the training for bootstrapping loop 2. Likely because of these samples, the dual-gate method, albeit still performing worse than the single-gate one, manages to gain some ground. With regard to the absolute figures of true positives returned (as opposed to accuracy), the two methods are even closer together after bootstrapping loop 1, whereas at that first loop, the single-gate method clearly outperforms the dual-gate one also on this measure.

Second, we note that the resulting dataset, containing 937 samples from the five normative categories, is not perfectly balanced. As table 7 in

[6] See here, last consulted on September 10, 2022.

the appendix, section C shows, the smallest sample size is in the deontological category with 137, while there are 301 samples in the Procedural category. Given our bootstrapping procedure, it has been impossible to achieve perfectly balanced sets without having to cut many good samples from the datasets.

Third, we suggest that, at this point, the results give us much reason to be optimistic. Using our bootstrapping process, we have been able to collect a dataset that is large enough and of sufficiently high quality to be useful to the community in many further applications. This in turn shows that the embeddings provided by pre-trained generic language models can provide enough information to build such a normative classifier. For instance, consider example (4), which the single-gate SVM of bootstrapping loop 3 has correctly classified as Rawlsian.

(4) Only a tax system that burdens exclusively the poorest group would be foreclosed on account of the difference principle, because that scheme of public finance would necessarily entail some redistribution, in the form of public goods at least, from the worst-off to the better-off.

What is remarkable about this correct prediction is that the typical superficial clues for Rawlsianism are all absent: mentioning "Rawls", emphasizing unjustifiable inequalities, etc. Rather, this sentence considers what taxation structures a central Rawlsian principle, namely the difference principle, excludes (rather than recommends).

6.3 Experiment 3: Annotation by Uninvolved Experts

We emphasize three insights provided by the results of this third experiment. First, the results support the reliability of the outcome of experiment 2. The fact that, in 85% of cases, one of the external annotators classified the samples in the same way as in the dataset suggests that, by and large, these classifications are reliable (Cohen's Kappa for this internal-external 2/3-agreement is at 0.81, see the appendix, section D).

Second, the classification is controversial, i.e., difficult. Annotators 1 and 2 diverge on their amount of agreement with annotator 0 (our internal annotator) by 19 percentage points, total agreement of all three annotators exists in only 51% of all cases. Likely, some of this divergence could be settled by discussing the samples in-person, but it still shows that this is a more complicated and controversial task than typical word-sense disambiguation. For instance, consider the example (5). Do you think it expresses a Deontological view, as it emphasizes equality of all individuals? While annotator 0 thought so, annotator 1 chose Utilitarian, probably because the sentence also suggests to focus on the (potential) welfare of everybody, that is, of the entire population. Thirdly, as annotator 3 did, you could also classify this sentence as Rawlsian, because it is about removing unjustified inequalities, namely such that concern an individual's potential to welfare.

(5) Social institutions should be designed to equalize the potential welfare of every individual.

Third, the variation between the normative categories is limited, not exceeding 18 percentage points. Given that the external annotators have not been involved in the specification of these categories (they were solely given the instructions that can be consulted in the appendix, section D), this gives reason to believe that these categories are sensible and hence useful to the community beyond the research lab that developed them.

7 Conclusion

In this article, we have explored the promises of using well-known classifying approaches together with state-of-the-art transformer-based PLMs to classify normative statements in the legal domain. Our results indicate that this approach does indeed hold substantial promise, which we would like to expand on in future research. In the meantime, we hope that our dataset will foster further research on this important, yet mostly uncharted, topic.

References

Elliott Ash, Malka Guillot, and Luyang Han. 2021. Machine extraction of tax laws from legislative texts. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 76–85, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual* *workshop on Computational learning theory*, pages 144–152.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198.

David Collier, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu. 2006. Essentially contested concepts: Debates and applications. *Journal of political ideologies*, 11(3):211–246.

Robert Dale. 2019. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Walter Bryce Gallie. 1955. Essentially contested concepts. In *Proceedings of the Aristotelian society*, volume 56, pages 167–198.

Clement Guitton, Aurelia Tamo-Larrieux, and Simon Mayer. 2022. A typology of automatically processable regulation. *Law, Innovation, and Technology*, 14(2).

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Alexandros Komninos and Suresh Manandhar. 2016. Dependency based embeddings for sentence classification tasks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1490–1500.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert Nozick. 1974. *Anarchy, state, and utopia*, volume 5038. new york: Basic Books.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jack W. Rae, Sebastian Borgeaud, and Trevor Cai et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. DeepMind Company Publication.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Edwina L Rissland and David B Skalak. 1989. Combining case-based and rule-based reasoning: A heuristic approach. In *IJCAI*, pages 524–530.

Philippe-André Rodriguez. 2015. Human dignity as an essentially contested concept. *Cambridge Review of International Affairs*, 28(4):743–756.

Kathryn E Sanders. 1991. Representing and reasoning about open-textured predicates. In *Proceedings of the 3rd international conference on Artificial intelligence and law*, pages 137–144.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jerrold Soh, How Khang Lim, and Ian Ernst Chai. 2019. Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Li Tang and Simon Clematide. 2021. Searching for legal documents at paragraph level: Automating label generation and use of an extended attention mask for boosting neural models of semantic similarity. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 114–122, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Matthijs J Warrens. 2015. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5(4):1.

Ludwig Wittgenstein. 2006/1953. Philosophische untersuchungen. In *Werkausgabe Band 1*. Suhrkamp.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Marco Wrzalik and Dirk Krechel. 2021. GerDaLIR: A German dataset for legal information retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Details on the dataset used

A.1 Sources for Experiment 1 and to Train for Bootstrapping Loop 1

Alm, J. & Melnik, M. I. (2005). Taxing the "Familiy" in the Individual Income Tax. Public Finance & Management, 5(1), 67-109.

Appelbaum, E. & Batt, R. (2017). Private Equity Partners Get Rich at Taxpayers Expense. Center for Economic and Policy Research.

Armstrong, C. (2013). Natural Resources: The Demands of Equality. Journal of Social Philosophy.

Avi-Yonah, R. S. (2002). Why Tax the Rich? Efficiency, Equity, and Progressive Taxation [Review of Does Atlas Shrug? The Economic Consequences of Taxing the Rich, by J. B. Slemrod]. The Yale Law Journal, 111(6), 1391–1416. https://doi.org/10.2307/797614

Barker, W. (2006). The Three Faces of Equality: Constitutional Requirements in Taxation. Case Western Reserve Law Review, 57(1), 1-53.

Baron, R. (2012). The Ethics of Taxation. Philosophy Now, 90.

Bezhanyan, R. (2017). Utilitarianism and Tax Policies.

Bird-Pollan, J. (2013). Unseating Privilege: Rawls, Equality of Opportunity, and Wealth Transfer Taxation. Wayne Law Review, 59(2), 713-742.

Bird-Pollan, J. (2016). Utilitarianism and Wealth Transfer Taxation. In Taxation of Wealth Transfers: A Philosophical Analysis, 124–151.

Bourguignon, F. (2018). Spreading the Wealth. Finance & Development, 55(1), 22-24.

Burgis, B. (2020). How to Debate Libertarians on Taxes — And Destroy Them. Jacobin.

Byrne, D. M. (1995). Progressive Taxation Revisited. Arizona Law Review, 37(3), 739-790.

Carens, J. H. (1986). Rights and Duties in an Egalitarian Society. Political Theory, 14(1), 31-50.

Cohn, A., Jessen, L. J., Klasnja, M. & Smeets, P. (2019). Why Do the Rich Oppose Redistribution? An Experiment with America's Top 5

Cooper, G. S. (1986). Income Tax Law and Contributive Justice: Some Thoughts on Defining and Expressing Consistent Theory of Tax Justice and Its Limitations. Australian Tax Forum, 3(3), 297-332.

Dodge, J. M. (2005). Theories of Tax Justice: Ruminations on the Benefit, Partnership, and Ability-to-Pay Principles. Tax Law Review, 58(4), 399-462.

Duff, D. G. (1993). Taxing Inherited Wealth: Philosophical Argument. Canadian Journal of Law and Jurisprudence, 6(1), 3-62.

Duff, D. G. (2005). Private Property and Tax Policy in a Libertarian World: A Critical Review. Canadian Journal of Law and Jurisprudence, 18(1), 23-45.

Durankev, B. (2019). Taxation and Social Justice. arXiv: General Economics. https://doi.org/10.48550/arXiv.1910.04155

Edwards, J. R. (2001). Taxation, Forced Labor, and Theft: Comment. The Independent Review, 6(2), 253–257.

Elkins, D. (2006). Horizontal Equity as Principle of Tax Theory. Yale Law & Policy Review, 24(1), 43-90.

Elkins, D. (2009). Taxation and the Terms of Justice. University of Toledo Law Review, 41(1), 73-106.

Epstein, R. A. (2005). Taxation with Representation: Or, the Libertarian Dilemma. Canadian Journal of Law and Jurisprudence, 18(1), 7-22.

Fleischer, M. (2010). Theorizing the Charitable Tax Subsidies: The Role of Distributive Justice. Washington University Law Review, 87(3), 505-566.

Frecknall-Hughes, J., Moizer, P., Doyle, E. & Summers, B. (2017). An Examination of Ethical Influences on the Work of Tax Practitioners. J Bus Ethics, 146, 729–745. https://doi.org/10.1007/s10551-016-3037-6

Galle, B. (2008). Tax Fairness. Washington and Lee Law Review, 65(4), 1323-1380.

Green, R. M. (1984). Ethics and Taxation: A Theoretical Framework. The Journal of Religious Ethics, 12(2), 146–161.

Gribnau, H. & Hughes-Frecknall, J. (2021). The Enlightenment and Influence of Social Contract Theory on Taxation. http://dx.doi.org/10.2139/ssrn.3963285

Hackney, P. (2021). Political Justice and Tax Policy: The Social Welfare Organization Case. Texas A&M Law Review, 8(2), 271-330. https://doi.org/10.37419/LR.V8.I2.2

Halliday, D. & Stewart, M. (2021). On "Dynastic" Inequality. In S. Gardiner (ed.) The Oxford Handbook of Intergenerational Ethics, 903.

Hänni, P. (2021). Chapter 9: The Swiss Tax System – Between Equality and Diversity. In The Principle of Equality in Diverse States, 253-289. https://doi.org/10.1163/9789004394612_011

Huemer, M. (2017). Is Taxation Theft? Libertarianism.org.

Hümbelin, O. & Farys, R. (2018). Income Redistribution Through Taxation – How Deductions Undermine the Effect of Taxes. Journal of Income Distribution, 25(1), 1-35.

Jestl, S. (2018). Inheritance Tax Regimes: A Comparison. The Vienna Institute for International Economic Studies. https://doi.org/10.3326/pse.45.3.3

Kamin, D. (2008). What Is Progressive Tax Change: Unmasking Hidden Values in Distributional Debates. New York University Law Review, 83(1), 241-292.

Kornhauser, M. E. (1995). Equality, Liberty, and Fair Income Tax. Fordham Urban Law Journal, 23(3), 607-662.

Lambert, P. J. & Naughton, H. T. (2009). The Equal Absolute Sacrifice Principle Revisited. Journal of Economic Surveys, 23(2), 328-349. http://dx.doi.org/10.1111/j.1467-6419.2008.00564.x

Leviner, S. (2006). From Deontology to Practical Application: The Vision of Good Society and the Tax System. Virginia Tax Review, 26(2), 405-446.

Leviner, S. (2012). The Normative Underpinnings of Taxation. Nevada Law Journal, 13(1), 95-133.

Lindsay, I. K. (2016). Tax Fairness by Convention: A Defense of Horizontal Equity. Florida Tax Review, 19(2), 79-119.

Mack, E. (2006). Non-Absolute Rights and Libertarian Taxation. Social Philosophy and Policy, 23(2), 109-141. https://doi.org/10.1017/S0265052506060195

Maloney, M. A. (1988). Distributive Justice: That is the Wealth Tax Issue. Ottawa Law Review, 20(3), 601-636.

Mankiw, N. G., Weinzierl, M. & Yagan, D. (2009). Optimal Taxation in Theory and Practice. Journal of Economic Perspectives, 23(4), 147-174. https://doi.org/10.1257/jep.23.4.147

McDaniel, P. R., & Repetti, J. R. (1993). Horizontal and Vertical Equity: The Musgrave/Kaplow Exchange. Florida Tax Review, 1(10), 607-622.

McIntyre, M. J. (1987). Tax Justice for Family Members after New York State Tax Reform. Albany Law Review, 51(3-4), 789-816.

Michael, M. A. (1997). Redistributive Taxation, Self-Ownership and the Fruit of Labour. Journal of

Applied Philosophy, 14(2), 137–146.

Milin, Z. (2014). Global Tax Justice and the Resource Curse: What Do Corporations Owe? Moral Philosophy and Politics, 1(1), 17-36. https://doi.org/10.1515/mopp-2013-0012

Miller, J. A. (2000). Equal Taxation: A Commentary. Hofstra Law Review, 29(2), 529-546.

Niesiobędzka, M., & Kołodziej, S. (2020). The Fair Process Effect in Taxation: The Roles of Procedural Fairness, Outcome Favorability and Outcome Fairness in the Acceptance of Tax Authority Decisions. Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues, 39(1), 246–253. https://doi.org/10.1007/s12144-017-9762-x

Ooi, V. (2016). Redistributive Taxation in the Modern World. Singapore Law Review, 34, 173-218.

Ozawa, M. N. (1973). Taxation and Social Welfare. Social Work, 18(3), 66–76.

Pawa, K. & Gee, C. (2021). Taxation and Distributive Justice in Singapore. IPS Working Papers, 42. Piketty, T. (2015). Capital, Inequality and Justice: Reflections on Capital in the Twenty-First Century. Basic Income Studies, 10(1), 141-156. https://doi.org/10.1515/bis-2015-0014

Porcano, T. M. (1984). Distributive Justice and Tax Policy. The Accounting Review, 59(4), 619–636.

Pressman, M. (2018). 'The Ability to Pay' in Tax Law: Clarifying the Concept's Egalitarian and Utilitarian Justifications and the Interactions between the Two. N.Y.U. Journal of Legislation & Public Policy, 21, 141-201.

Pryor, A. (2011). Ought There to Be Graduated Federal Income Tax: Is Robin Hood Justice, Justice at All? Georgetown Journal of Law & Public Policy, 9(2), 543-562.

Scheuer, F. (2020). Taxing the Superrich: Challenges of a Fair Tax System. UBS Center Public Paper.

Simester, A. A., & Chan, W. (2003). On Tax and Justice. Oxford Journal of Legal Studies, 23(4), 711-726.

Slemrod, J. (1998). The Economics of Taxing the Rich, National Bureau of Economic Research, 1-38. https://doi.org/10.3386/w6584

Stark, J. (2022). Tax Justice Beyond National Borders—International or Interpersonal? Oxford Journal of Legal Studies, 42(1), 133-160. https://doi.org/10.1093/ojls/gqab026

Steuerle, C. E. (2002). An Equal (Tax) Justice for All. Tax Justice, 253-284.

Sugin, L. (2004). Theories of Distributive Justice and Limitations on Taxation: What Rawls Demands from Tax Systems. Fordham Law Review, 72(5), 1991-2014.

Sugin, L. (2016). Rhetoric and Reality in the Tax Law of Charity. Fordham Law Review, 84(6), 2607-2632.

Svoboda, V. (2016). Libertarianism, Slavery, and Just Taxation. Humanomics: The International Journal of Systems and Ethics, 32(1), 69-79. https://doi.org/10.1108/H-05-2015-0031

Taite, P. C. (2014). Exploding Wealth Inequalities: Does Tax Policy Promote Social Justice or Social Injustice. Western New England Law Review, 36(3), 201-220.

Vallentyne, P. (2018). Libertarianism and Taxation. In M. O'Neill & S. Orr (ed.), Taxation: Philosophical Perspectives, 98–110.

Young, H. P. (1987). Progressive Taxation and the Equal Sacrifice Principle. Journal of Public Economics, 32, 203-214.

A.2 Sources for Predictions in the first Bootstrapping Loop

Avi-Yonah, R., Avi-Yonah, O., Fishbien, N., & Xu, H. (2020). Federalizing Tax Justice. Indiana Law Review, 53(3), 461-498. http://dx.doi.org/10.2139/ssrn.3249010

Baird, C. W. (1981). Proportionality, Justice, and the Value-Added Tax. Cato Journal, 1(2), 405-420.

Barker W. (2005). Expanding the Study of Comparative Tax Law to Promote Democratic Policy: The Example of the Move to Capital Gains Taxation in Post-Apartheid South Africa. Penn State Law Review, 109(3), 703-727.

Crawford, P. (2014). Occupy Wall Street, Distributive Justice, and Tax Scholarship: An Ideology Critique of the Consumption Tax Debate. University of New Hampshire Law Review, 12(2), 137-174.

Dagan, T. (2017). International Tax and Global Justice. Theoretical Inquiries in Law, 18(1), 1-36. http://dx.doi.org/10.2139/ssrn.2762110

Flynn, J. J., & Ruffinengo, P. (1975). Distributive Justice: Some Institutional Implication of Rawls' Theory of Justice. Utah Law Review, 1975(1), 123-157.

Fried, B. H. (1999). The Puzzling Case for Proportionate Taxation. Chapman Law Review, 2, 157-

196.

Kamin, D. (2008). What Is Progressive Tax Change: Unmasking Hidden Values in Distributional Debates. New York University Law Review, 83(1), 241-292.

Kaplow, L. (2007). Discounting Dollars, Discounting Lives: Intergenerational Distributive Justice and Efficiency. University of Chicago Law Review, 74(1), 79-118.

Kenealy, W. J. (1961). Equal Justice under Law Tax Aid to Education. Catholic Lawyer, 7(3), 183-202.

Kurland, N. G. (1977). Beyond ESOP: Steps Toward Tax Justice–Part 2. Tax Executive, 29(4), 386-402.

Maier, C. & Schanz, D. (2017). Towards Neutral Distribution Taxes and Vanishing Tax Effects in the European Union. http://dx.doi.org/10.2139/ssrn.2948475

McIntyre, M. J. (1988). Implications of US Tax Reform for Distributive Justice. Australian Tax Forum, 5(2), 219-256.

Murphy, L. B. (1996). Liberty, Equality, Well-Being: Rakowski on Wealth Transfer Taxation. Tax Law Review, 51(3), 473-494.

Repetti, J., & Ring, D. (2012). Horizontal Equity Revisited. Florida Tax Review, 13(3), 135-156.

van Apeldoorn, L. (2019). A Sceptic's Guide to Justice in International Tax Policy. Canadian Journal of Law and Jurisprudence, 32(2), 499-512. https://doi.org/10.1017/cjlj.2019.14

A.3 Sources for Predictions in the Second Bootstrapping Loop

Banfi, T. (2015). A Fair Tax (System) or an Ethical Taxpayer? Society and Economy, 37, 107–116. https://doi.org/10.1556/204.2015.37.s.7

Bărbuţă-Mişu, N. (2011). A Review of Factors for Tax Compliance. Economics and Applied Informatics, 1, 69-76.

Bradley, B., & Gephardt, R. (1984). Fixing the Income Tax with the Fair Tax. Yale Law & Policy Review, 3(1), 41–57.

Braithwaite, V. (2003), Who's Not Paying their Fair Share: Public Perceptions of the Australian Tax System. Australian Journal of Social Issues, 38, 323-348. https://doi.org/10.1002/j.1839-4655.2003.tb01149.x

Braithwaite, V. (2003). Tax System Integrity and Compliance: The Democratic Management of the Tax System. In Taxing Democracy, 269–287.

DeConcini, D. (1985). A Proposed Simple and Fair Tax. Journal of Legislation, 12(2), 143-155.

Gephardt, R. A., & Bryant, E. G. (1985). The Fair Tax Act: A Plan for Simple, Fair, and Economically Rational Tax. Journal of Legislation, 12(2), 129-142.

Herzberg, A. (1963). Blueprint of Fair Tax Administration. Taxes - The Tax Magazine, 41(3), 161-164.

McKerchar, M. (2008). Philosophical Paradigms, Inquiry Strategies and Knowledge Claims: Applying the Principles of Research Design and Conduct to Taxation. eJournal of Tax Research, 6(1), 5-22.

Osberg, L. (2016). 2. What's Fair?: The Problem of Equity in Taxation. In A. Maslove (Ed.), Fairness in Taxation, 63-86. https://doi.org/10.3138/9781442623293-004

Rosow, S. (1984). Treasury's Tax Reform Proposals: Not A Fair Tax. Yale Law & Policy Review, 3(1), 58-72.

Schoenblum, J. A. (1995). Tax Fairness or Unfairness A Consideration of the Philosophical Bases for Unequal Taxation of Individuals. American Journal of Tax Policy, 12(2), 221-272.

Slemrod, J. (2002). Tax Systems. The Reporter, 3, 1-17.

Slemrod, J. (2018). Is This Tax Reform, or Just Confusion? Journal of Economic Perspectives, 32(4), 73-96. https://doi.org/10.1257/jep.32.4.73

Sugin, L. (2011). A Philosophical Objection to the Optimal Tax Model. Tax Law Review, 64(2), 229-282.

A.4 Sources for Predictions in the Third boostrapping loop

Bello, K.B., & Danjuma, I.M. (2014). Review of Models/Theories Explaining Tax Compliance Behavior. Sains Humanika, 2(3), 35-38. https://doi.org/10.11113/sh.v2n3.432

Benham, F. (1942). What is the Best Tax-System? Economica, 9(34), 115–126. https://doi.org/10.2307/2549805

Bogenschneider, B. (2017). A Philosophy Toolkit for Tax Lawyers. Akron Law Review, 50(3), 452-494.

Buehler, A. G. (1949). The Cost and Benefit Theories. Tax Law Review, 5(1), 17-34.

Cobham, A. (2005). Taxation Policy and Development. OCGG Economy Analysis, 2, 1-23.

Colm, G. (1934). The Ideal Tax System. Social Research, 1(3), 319–342.

Colm, G. (1940). Conflicting Theories of Corporate Income Taxation. Law and Contemporary Problems, 7(2), 281-290.

Diamond, P. A., & Mirrlees, J. A. (1971). Optimal Taxation and Public Production I: Production Efficiency. The American Economic Review, 61(1), 8–27.

Dimopoulos, T. (2015). Theories and Philosophy of Property Taxation.

Dom, R. & Miller, M. (2018). Reforming tax systems in the developing world: What can we learn from the past? Overseas Development Institute (ODI).

Dorocak, J. R. (2015). What Would Libertarian Tax Look Like. South Texas Law Review, 57(2), 147-168.

Escarraz, D. R. (1967). Wicksell and Lindahl: Theories of Public Expenditure and Tax Justice Reconsidered. National Tax Journal, 20(2), 137–148.

Fagan, E. D. (1938). Recent and Contemporary Theories of Progressive Taxation. Journal of Political Economy, 46(4), 457–498.

Fausto, D. (2008). The Italian theories of progressive taxation. The European Journal of the History of Economic Thought, 15(2), 293-315. https://doi.org/10.1080/09672560802037607

Feser, E. (2000). Taxation, Forced Labor, and Theft. The Independent Review, 5(2), 219–235.

Hamlin, A. (2018). What Political Philosophy Should Learn from Economics about Taxation. In M. O'Neill & S. Orr (ed.), Taxation: Philosophical Perspectives, S. 1–28.

Hassett, K.& Auerbach A. (2005). Toward Fundamental Tax Reform, American Enterprise Institute.

Hodgson, H. (2010). Theories of Distributive Justice: Frameworks for Equity. Journal of the Australasian Tax Teachers Association, 5, 86-116.

Horvitz, J. S. (1977). Theories of Legal Responsibility in Regard to the CPA in Tax Practice. Baylor Law Review, 29(3), 475-498.

Howard, J. M. (1992). When Two Tax Theories Collide: A Look at the History and Future of Progressive and Proportionate Personal Income Taxation. Washburn Law Journal, 32(1), 43-76.

Josheski, D. & Boshkov T. (2020). Critical Review of the (Second Wave) Optimal Tax Theories. University Goce Delcev-Shtip. http://dx.doi.org/10.2139/ssrn.3531287

Kiser, E. (1994). Markets and Hierarchies in Early Modern Tax Systems: A Principal-Agent Analysis. Politics & Society, 22(3), 284–315. https://doi.org/10.1177/0032329294022003003

Kordana, K., & Tabachnick, D. (2006). Taxation, the Private Law, and Distributive Justice. Social Philosophy and Policy, 23(2), 142-165. https://doi.org/10.1017/S0265052506060201

LeFevre, T. A. (2017). Justice in Taxation. Vermont Law Review, 41(4), 763-798.

McCaffery, E. J. (1994). The Political Liberal Case Against the Estate Tax. Philosophy & Public Affairs, 23(4), 281–312.

McCaffery, E. J., & Hines, J. (2010). The Last Best Hope for Progressivity in Tax. Southern California Law Review, 83(5), 1031-1098.

Misra, F. (2019). Tax Compliance: Theories, Research Development and Tax Enforcement Models. Accounting Research Journal of Sutaatmadja, 3(2), 189-204. https://doi.org/10.35310/accruals.v3i2.72

Muzurura, J., Nyoni, J. & Mataruka, L. (2021). The Anatomy of Tax Evasion and Tax Morale: Lessons from Tax Theories, Tax Audits and Surveys in Zimbabwe. International Journal of Social Science and Economic Research, 6(4), 1283- 1303. https://doi.org/10.46609/IJSSER.2021.v06i04.011

Panova, T. V. & Panov, E. G. (2021). Tax philosophy versus fiscal sociology: choice problem in teaching, SHS Web of Conferences, 103, 1-4. https://doi.org/10.1051/shsconf/202110301027

Pirttila, J. (1999). Tax Evasion and Economies in Transition: Lessons from Tax Theory. BOFIT Discussion Paper, 2. http://dx.doi.org/10.2139/ssrn.1016663

Saez, E. & Stantcheva, S. (2016). Generalized Social Marginal Welfare Weights for Optimal Tax Theory. American Economic Review, 106(01), 24-45. https://doi.org/10.3386/w18835

Sahota, G. S. (1978). Theories of Personal Income Distribution: A Survey. Journal of Economic Literature, 16(1), 1–55.

Salahuddin, A. (2018). Robert Nozick's Entitlement Theory of Justice, Libertarian Rights and the Minimal State: A Critical Evaluation. Journal of Civil & Legal Sciences, 7(1), 234-238. https://doi.org/10.4172/2169-0170.1000234

Samuelson, P. A. (1958). Aspects of Public Expenditure Theories. The Review of Economics and Statistics, 40(4), 332–338. https://doi.org/10.2307/1926336

Slemrod, J. (2022). Group Equity and Implicit

Discrimination in Tax Systems. National Tax Journal, 75(1). https://doi.org/10.1086/717960

Sugin, L. (2004). Theories of Distributive Justice and Limitations on Taxation: What Rawls Demands from Tax Systems. Fordham Law Review, 72(5), 1991-2014.

Sunderman, M., Birch, J., Cannaday, R. & Hamilton, T. (1990). Testing for Vertical Inequity in Property Tax Systems, Journal of Real Estate Research, 5(3), 319-334, https://doi.org/10.1080/10835547.1990.12090625

van der Vossen, Bas. (2017). Libertarianism. Oxford Research Encyclopedia of Politics.

B Details on Experiment 1

SVM Hyperparameters & Implementation Details We use the following different hyperparameters for our search:[7]

C Regularization parameter, inversely proportional to strength of regularization – a large C causes individual training samples to influence the resulting function stronger: 0.1, 1, 10, 100, 1000

kernel Kind of kernel used in the SVM: rbf (radial basis function), poly (polynomial), linear

gamma Specifies the sphere of influence of datapoints on the resulting SVM: 1, 0.1, 0.01, 0.001, 0.0001

We have used scikit-learn's default implementation of SVM that automatically chooses one-vs.one for classification tasks with more than two classes, and it automatically employs five-fold cross-validation.

Models & Embedding Types We are testing three different kinds of models; for references, see above, section 2; for the full list of models, see below, table 5. We use four different routines to extract the embeddings:

Sentence-Averaged Word-Based In this routine, we use the average of all word embeddings, as the model delivers it for all words in the sentence. Hence, the sentence-embedding used here is the average of all word embeddings whose words appear in the sentence. Here, we use well-researched transformer-based PLMs, namely RoBERTa and BERT, but also models fine-tuned to first-order legal domains such as legal-bert (see above, section 2)

Sentence-based Here, we use the embeddings, as they directly result from the sentence-bert models trained by Reimers and Gurevych 2019. These models also output the average of all word embeddings (which we manually compute in the second variant), but they have been fine-tuned on the sentence level by training them on a wide variety of sentence-level tasks and datasets (the original models reported in Reimers and Gurevych 2019 use the combination of the SNLI and the Multi-Genre NLI datasets). Furthermore, the models that they fine-tuning are of many flavors, ranging from classical BERT to recent proposals such as mpnet (see above, section 2).

Average of Classical Word Embeddings We here test two classical kinds of word embeddings, GloVE as well as Komninos (see above, section 2), again taking the average of all word embeddings as the sentence embedding.

Table 5 lists all of the models used.

Word-Based Models
bert-base-cased
bert-large-cased
roberta-large
legal-bert-base-uncased
SBERT-Models
paraphrase-TinyBERT-L6-v2
paraphrase-distilroberta-base-v2
paraphrase-mpnet-base-v2
paraphrase-multilingual-mpnet-base-v2
paraphrase-MiniLM-L12-v2
paraphrase-MiniLM-L6-v2
paraphrase-albert-small-v2
paraphrase-multilingual-MiniLM-L12-v2
paraphrase-MiniLM-L3-v2
nli-mpnet-base-v2
nli-roberta-base-v2
nli-distilroberta-base-v2
distiluse-base-multilingual-cased-v1
stsb-mpnet-base-v2
stsb-distilroberta-base-v2
distiluse-base-multilingual-cased-v2
stsb-roberta-base-v2
Classical Models
average_word_embeddings_glove.6B.300d
average_word_embeddings_komninos

Table 5: Overview on the 23 models tested In clustering.

Table 6 lists all models whose embedding were used in experiment 1 with the accuracies of the best performing SVM that was found in the hyperparameter search specifically for these embeddings. For instance, The embeddings of roberta-large can be

[7] Compare the details here, last consulted on September 16, 2022.

Modelname	Type	Norm	5cat
pp-ml-mpnet-base-v2	sbert	91%	87%
pp-mpnet-base-v2	sbert	91%	85%
nli-mpnet-base-v2	sbert	91%	84%
stsb-roberta-base-v2	sbert	94%	83%
stsb-droberta-base-v2	sbert	94%	83%
nli-droberta-base-v2	sbert	95%	82%
roberta-large	avword	98%	82%
nli-roberta-base-v2	sbert	94%	82%
pp-droberta-base-v2	sbert	95%	80%
stsb-mpnet-base-v2	sbert	91%	80%
du-base-ml-cased-v2	sbert	91%	77%
pp-MiniLM-L12-v2	sbert	91%	77%
pp-MiniLM-L6-v2	sbert	91%	77%
du-base-ml-cased-v1	sbert	91%	77%
awe_komninos	sbert	91%	77%
bert-large-cased	avword	91%	77%
pp-ml-MiniLM-L12-v2	sbert	91%	77%
bert-base-cased	avword	91%	76%
pp-TinyBERT-L6-v2	sbert	91%	76%
awe_glove.6B.300d	sbert	91%	75%
pp-MiniLM-L3-v2	sbert	91%	75%
pp-albert-small-v2	sbert	91%	74%
nlpaueb-legalbertbase	avword	91%	74%

Table 6: Results of classifying samples as belonging to one of the five normative categories (35 samples each, column 5cats) and as normative or nonnormative (175/1708 samples, column Norm). Most frequent class baseline reaches accuracy of 20% for 5cat and 91% for Norm. "pp" = paraphrase, "ml" = multilingual, "droberta" = distilroberta, "du" = distiluse, "awe" = average_word_embedding.

combined with an SVM to form a classifier that delivers 98% accuracy in the normative-nonnormative task and 82% at the 5cat task.

C Details on Experiment 2

Table 7 shows the distribution of samples across the normative classes in the final dataset that results from a combination of the corrected outputs from both methods after bootstrapping loop 3 with any duplicates removed.

D Details on Experiment 3

Category	# Samples
Deontological	137
Libertarian	159
Procedural	301
Rawlsian	138
Utilitarian	202
None	2194
Total Normative	937
Grand Total	3131

Table 7: Samples by category and in total in the final dataset, combining the reviewed output from bootstrapping loop 3 by both methods, and having removed any duplicates.

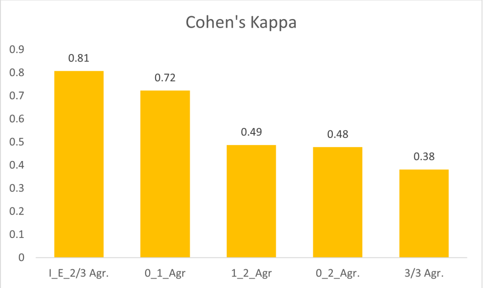


Figure 3: Cohen's Kappa for the inter-annotator agreement in experiment 3.

Figure 3 gives Cohen's Kappa for the agreement between our three annotators; briefly, Cohen's Kappa gives an inter-annotator agreement that takes into account the statistical probability of annotators agreeing by mere chance (see Warrens 2015 for further details). As can be seen, the basic layout doesn't change when compared to the accuracies reported above, figure 2: Internal-External-2/3-agreement is highest, annotator 2 diverges from 0 and 1 quite often, 3/3-agreement is lowest.

Table 8 gives the inter-annotator agreement by source of sample. For instance, the inter-annotator agreement with samples that were selected by our expert directly (as opposed to building on predictions by a classifier called "Fully human") is highest both in internal-external 2/3 agreement and 3/3 agreement. Table 8 shows that the origin of the samples does make a difference for the overall inter-annotator agreement, but a relatively small one, not exceeding 12 percentage points in the internal-external 2/3 agreement. This adds further evidence to the claim that our internal annotator has not been overly biased towards the output of our classifiers. Otherwise, we would expect annotators 1 and 2 to diverge from annotator 1 much more often regarding machine-produced samples than regarding

Origin	Count	I_E_2/3 Agr.	3/3 Agr.
Fully Human	175	90%	65%
Single-Gate	122	89%	51%
Dual-Gate	353	78%	41%

Table 8: Inter-annotator agreement by origin of the samples("I_E_2/3" continues to represent the 2/3-agreement where one of the agreeing annotators is our internal annotator 0, the other is either 1 or 2).

fully-human compiled samples.

In the remainder of this section, we give the literal instructions given to annotators, anonymized for reviewing.

General Task Description Thank you very much for taking the time to annotate our samples and thereby contribute to the ongoing NLP project. In the following, we provide instructions to ensure that your annotations are maximally useful to the project. Please read through the entire paper before annotating. Let me know if you have any questions: ANEMAIL.

For the list of statements enclosed, you are asked to make two decisions for each sample:

1. Decide whether the sample expresses a normative statement: If you think it does, enter "YES" into column A "Annotator Norm", if you think not, enter "NO". Please make sure you type it in all caps without any blanks.

2. If you have answered "YES" for a given sample, decide to which of the five normative categories the sample belongs; if you are unable to assign the sample to any of the five categories, use "OTHER"; please only use this category if you are fully convinced that the sentence does not fit any of the categories. Depending on your judgment, enter one of the following into Column B "Annotator Cat" (again, make sure you type it without blanks, and always in the exact way specified here):

   (a) Libertarian
   (b) Rawlsian
   (c) Deontological
   (d) Procedural
   (e) Utilitarian
   (f) OTHER

Details on categorization

1. **Normative vs. Not Normative**: Does the statement (a) make a direct recommendation what the state, an individual, etc. should be doing, or (b) does the statement make an assertion about what is just/unjust, fair/unfair, moral/immoral? If either (a) or (b) applies, the statement is normative.
   Examples:

   (a) **Not normative**: "An income tax can be used to redistribute taxable income."
   (b) **Normative**: "All that matters for the Utilitarian is maximizing utility, and by distributing the tax cut across income classes, a previously optimal tax system would no longer be so."

2. Following is a brief description of the normative categories that can help you decide about categorization. We are aware that the categorization proposed here is not beyond dispute; for the present project, we ask you to simply adhere to the categorization sketched here. Let us know if any of the categories were particularly challenging during the annotation process.

   (a) **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. Therefore, taxation should be kept at an absolute minimum, what is needed to ensure that a minimal state is functioning.
   Examples:
      i. Nozick likens the imposition of redistributive taxes (typically progressively designed) on people who are working to earn money to partial enslavement.
      ii. the anti-progressive tax argument is often characterized as an argument that every person has a responsibility to take care of himself, and no one, including the wealthy, has an obligation to assist those in need.

   (b) **Rawlsians** in contrast, hold that the state should redistribute wealth and income to the extent to which this can reduce unjustified inequalities in the distribution of wealth. Rawlsians hold that many inequalities are in fact unjust, including,

for instance, the wealth of the family into which one is born, or the quality of the schools that are available in your area. As a consequence, Rawlsians will typically defend progressive taxation of both income and wealth.
Examples:

i. The increasing inequality of market income can be significantly ameliorated by the redistributive effect of the tax transfer system, if it is appropriately targeted.
ii. By distributing the tax burden more onerously on those who have the most physical wealth, equality of opportunity goals will be furthered.

(c) The term **Deontological** ethics covers a broad variety of positions. For the purpose of the present annotations, we consider positions as Deontological if they focus on the treatment of the individual 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).
Examples:

i. Max burdens should bear similarly upon persons whom we regard as in substantially similar circumstances, and differently where circumstances differ.
ii. Horizontal equity requires equals to be treated equally

(d) **Procedural** positions hold 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 aimed at achieving a just societal structure based on a democratic decision-making process.
Examples:

i. For Locke himself, the key institutional requirement was that taxes should not be levied except by "the consent of the people," which he understood as "the consent of the majority, giving it either by themselves, or their representatives chosen by them."
ii. As expected, respondents were more accepting of changes introduced in a fair manner than in an unfair manner, even if the changes resulted in higher tax burdens.

(e) **Utilitarian** positions emphasize the effect on overall happiness or welfare that a certain tax provision has. Hence, rather than capitalizing on participative, democratic decision-making, the equal treatment of individuals, or reducing unjustified inequalities, Utilitarians consider the overall net increase or decrease in wealth, happiness, or welfare, that a tax provision has on the society in question. Often, Utilitarians argue that the least well-off should benefit most from redistribution caused by taxation because their happiness shows the largest relative increase if they receive a certain amount of money.
Examples:

i. Efficiency analysis looks to overall social welfare as a measure of a tax's virtue.
ii. 68 Inequality is considered unfair because of the arbitrariness of unequal outcomes.69 But this inequality can potentially be justified in fairness terms if those at the bottom are made better off because of it.