Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210315

Generation of Legal Norm Chains: Extracting the Most Relevant Norms from Court Rulings

Ingo GLASER [a,1], Sebastian MOSER [a] and Florian MATTHES [a]
[a] *Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany*

Abstract. Various online databases exist to make judgments accessible in the digital age. Before a legal practitioner can utilize state-of-the-art information retrieval features to retrieve relevant court rulings, the textual document must be processed. More importantly, many verdicts lack crucial semantic information which can be utilized within the search process. One piece of information that is frequently missed, as the judge is not adding it during the publication process within the court, is the so-called norm chain. This list contains the most relevant norms for the underlying decision.

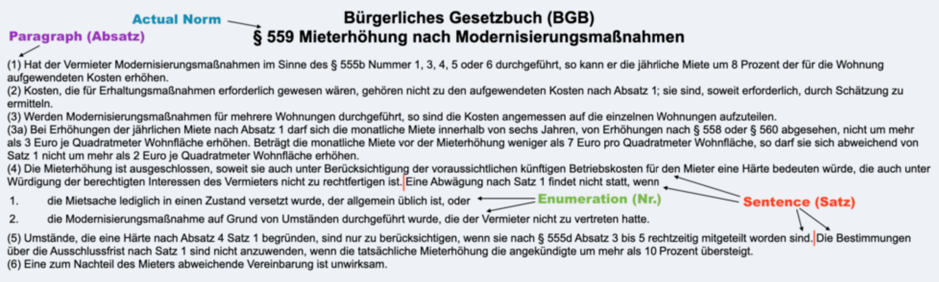
Therefore this paper investigates the feasibility of automatically extracting the most relevant norms of a court ruling. A dataset constituting over 42k labeled court rulings was used in order to train different classifiers. While our models provide F1 performances of up to 0.77, they can undoubtedly be utilized within the editorial publication process to provide helpful suggestions.

Keywords. natural legal language processing, norm chains, legal court rulings, multi-label classification

1. Introduction

Legal research constitutes a significant part of the daily work of a legal practitioner, particularly lawyers [1,2]. As the work of legal workers is not just knowledge-driven but also time-consuming, recent research activities and the industry try to support the legal research process. The focus here is often on legal information retrieval. That is why various online databases exist that provide convenient search functionalities. However, the path from a textual court ruling (in the remainder of this paper the terms ”verdict”, ”ruling”, and ”decision” all refer to the entire court ruling document), as it is created by a judge, into an online database is often tedious and involves a vast amount of human labor. This path includes typical processing tasks such as segmentation or information extraction and enrichment with semantic information. One type of such semantic information is the so-called norm chain.

[1] Corresponding Author: Ingo Glaser, Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany; E-mail: ingo.glaser@tum.de



Bürgerliches Gesetzbuch (BGB)
§ 559 Mieterhöhung nach Modernisierungsmaßnahmen

(1) Hat der Vermieter Modernisierungsmaßnahmen im Sinne des § 555b Nummer 1, 3, 4, 5 oder 6 durchgeführt, so kann er die jährliche Miete um 8 Prozent der für die Wohnung aufgewendeten Kosten erhöhen.
(2) Kosten, die für Erhaltungsmaßnahmen erforderlich gewesen wären, gehören nicht zu den aufgewendeten Kosten nach Absatz 1; sie sind, soweit erforderlich, durch Schätzung zu ermitteln.
(3) Werden Modernisierungsmaßnahmen für mehrere Wohnungen durchgeführt, so sind die Kosten angemessen auf die einzelnen Wohnungen aufzuteilen.
(3a) Bei Erhöhungen der jährlichen Miete nach Absatz 1 darf sich die monatliche Miete innerhalb von sechs Jahren, von Erhöhungen nach § 558 oder § 560 abgesehen, nicht um mehr als 3 Euro je Quadratmeter Wohnfläche erhöhen. Beträgt die monatliche Miete vor der Mieterhöhung weniger als 7 Euro pro Quadratmeter Wohnfläche, so darf sie sich abweichend von Satz 1 nicht um mehr als 2 Euro je Quadratmeter Wohnfläche erhöhen.
(4) Die Mieterhöhung ist ausgeschlossen, soweit sie auch unter Berücksichtigung der voraussichtlichen künftigen Betriebskosten für den Mieter eine Härte bedeuten würde, die auch unter Würdigung der berechtigten Interessen des Vermieters nicht zu rechtfertigen ist. Eine Abwägung nach Satz 1 findet nicht statt, wenn
1. die Mietsache lediglich in einen Zustand versetzt wurde, der allgemein üblich ist, oder
2. die Modernisierungsmaßnahme auf Grund von Umständen durchgeführt wurde, die der Vermieter nicht zu vertreten hatte.

(5) Umstände, die eine Härte nach Absatz 4 Satz 1 begründen, sind nur zu berücksichtigen, wenn sie nach § 555d Absatz 3 bis 5 rechtzeitig mitgeteilt worden sind. Die Bestimmungen über die Ausschlussfrist nach Satz 1 sind nicht anzuwenden, wenn die tatsächliche Mieterhöhung die angekündigte um mehr als 10 Prozent übersteigt.
(6) Eine zum Nachteil des Mieters abweichende Vereinbarung ist unwirksam.

Figure 1. Annotated screenshot of BGB § 559 taken from www.gesetzte-im-internet.de

A norm chain is a series of legal norms that lead to the consequence of a verdict. Thus, with the help of legal norm chains, the practiced civil lawyers can save the time-consuming step of searching for the relevant and referenced legal norms. Therefore, a general definition of the legal norm chain is the following:

> The chain of norms is a combination of explicitly or implicitly referencing legal norms that extends from a legal consequence order to the lowest level of the facts.

Hereby, a legal norm is understood to be either a statutory regulation or a rule of a general abstract nature issued on a statutory basis or contained in the common law. Figure 1 provides an example of a norm from the German Civil Code (BGB). The norm would be referenced as ”BGB § 559”. The shown norm regulates the increase of rent after modernization measures. Depending on the context, a more granular reference can be made such as ”BGB § 559 Absatz 3”, which would refer to the paragraph after ”(3)”. Going one stop further, it is also possible to reference a precise sentence within a paragraph, e.g., ”BGB § 559 Absatz 4, Satz 1”. Last but not least, enumerations might be referenced as well, such as ”BGB § 559 Absatz 4 Satz 2 Nr. 1”. Now, a norm chain consists of one-to-many of such norm references of varying granularities. An example of such a norm chain would be ”BGB § 535 Abs. 1 Satz 2, § 536 Abs. 1, § 536a Abs. 2 Nr. 1”. This chain consists of three different norms. The first norm references a concrete sentence, the second norm constitutes a complete paragraph, and the final norm refers to a specific enumeration number. That example was taken from a verdict from the German Supreme Court (VIII ZR 271/17).

While particularly verdicts from higher courts, such as the German Supreme Court, usually contain the norm chain as the judge provides it, the vast majority of court rulings leave the court’s internal publication process without it. That is why legal authors participating in the editorial process of a legal publisher are required to create the norm chain afterward.

Therefore, in this work, we want to investigate the feasibility of extracting relevant norms to automatically create norm chains utilizing natural language processing (NLP). The remainder of this paper is structured as follows: Section 2 describes the data set utilized in this work together with required pre-processing steps. The applied methods are discussed in Section 3. Detailed analysis and discussion of our results are provided in Section 4. Limitations of the presented work, along with a short overview of related work, are provided in Section 5 before Section 6 closes with a conclusion and outlook.

2. Data

For this research, we used a dataset of 42k German court rulings from various instance levels. The verdicts were given to us by a German legal publisher via an annotated XML format that contains the norm chain separated into its containing norms and the whole text of the verdict. Norm references within the text are annotated. Based on this, we extracted the text from the different verdict sections and segmented it into sentences via spaCy[2]. The norms from the norm chain were extracted, cross-referenced against a list[3] with all existing norms to validate the classification targets, and then cleaned (more precisely, special characters were stripped away, and additional information such as dates were removed). For each verdict, we also extracted all the referenced norms from the text sections utilizing the annotations, and regular expressions as some norms are not annotated. The referenced norms were then also validated against the list from *gesetze-im-internet.de*.

Based on this preprocessing process, our dataset contains 8,359 different, unique classification targets (e.g., *BGB 3*) extracted from a total of 111k norms from all norm chains with an average number of 2.6 norms per norm chain. When only considering the specific norm without any paragraph or section reference (e.g., *BGB*), there are 666 different targets for classification with on average 1.8 norms per norm chain. We will call this the reduced target set, and when referencing a specific norm from this set, we will call it the reduced norm. Overall, the distribution of norms is highly skewed as, on average, a norm only appears in 3.6 norm chains. This number is slightly better for reduced norms as they appear on average in 11.7 norm chains, but their occurrences still follow a power-law distribution. Some norms appear more frequently, such as *BGB* or *ZPO*, but others are only found once in our dataset. Around 20% of the target norms are from 14 different norm texts, and around 3.500 norms are only used once. We cannot expect reliable results for norms that are not used very often. Any prediction score will be skewed by this imbalance, which we want to quantify during our analysis stage. Based on this, our problem can be described as multi-label classification as each verdict can have multiple important norms assigned, and we need to select those from a large set of possible targets.

When looking at the referenced norms, we see that only 55% of the norms in norm chains are actually cited within the verdict itself (with an exact match for paragraph, section, etc.). For the reduced norms, a much more significant percentage (94%) is found within the content of the court ruling. However, some norms are not referenced while still being identified as an essential norm towards that decision (i.e., they appear in the norm chain.

Lastly, we randomly selected 10% of the dataset as a test set for our final evaluation. From the remaining verdicts, we again used 10% for a development set to select the best hyperparameters during training of the different classifiers.

3. Methods

We applied four different classification models with varying complexity levels to tackle this classification problem. We used one linear layer for the classification on top of the

[2] spacy.io

[3] The list was extracted from gesetze-im-internet.de

Table 1. Micro-F1 test set performance for the each model type and each classification target.

Method	Reduced Norm	Norm
word2vec	23.10	8.86
TF/IDF	77.05	52.57
Ref. Norms	71.48	49.15
BERT	53.88	31.04

featurization described next. First, we used 100-dimensional word2vec embeddings [3] which were trained on a different legal corpus with around 50k sentences. This corpus contains many different German laws and court rulings to have a wider variety in terms of textual content. The embeddings were trained with Gensim [4] using a word window of 5 while unknown words are replaced with an all-zeros embedding. We then average all word embeddings in a court ruling. Second, we used the occurrence of a norm in the court ruling text as features for classification. The occurrences are one-hot encoded, i.e., if a specific norm is referenced within the decision, its corresponding entry in the feature vector is set to 1. We assume that it is possible to determine the most important norms for the norm chain based on the mixture of referenced norms. Our third model uses TF/IDF as a feature with a minimum document frequency of 50 per word. The final model we tested is utilizing *bert-base-german-cased* a BERT-model [5] which is pre-trained on a German corpus which also contains legal documents. As the number of usable words is limited for the BERT-based classification, we used the first 512 tokens of a court ruling.

All models were implemented in PyTorch[4]. We used Adam as the optimizer for the Binary-Cross-Entropy loss. To determine the best learning rate for each model type, we first did manual testing to discover an appropriate learning rate range. We then randomly sampled three learning rates from those ranges per type and classification target (norm vs. reduced norm) and selected the best model through the micro-F1 score on the development set. We decided to use this optimization scheme, as a grid-search would have been too costly, but more importantly, each model needs slightly different learning rates to obtain optimal results. We trained each model for 100 epochs with early stopping based on the development set micro-F1 score and a patience of 10 epochs, except for the BERT-based model. Here we only trained for 25 epochs as the model converged faster, and no significant improvements were observed afterward. The micro-F1 score for each model on the test set can be seen in Table 1 for the norm set and the reduced norm set.

4. Model Analysis

As seen in Table 1, the averaged word2vec featurization has the worst performance by a wide margin. For that reason, we do not further investigate this model.

Surprisingly, TF/IDF and the referenced norms outperform the BERT-based classifier by almost 20% for both target sets. A possible reason for the lower performance of BERT could be the text type. German court rulings are relatively long, and we needed to make some simplifying assumptions to encode the court ruling. We could not identify any evidence towards this claim for the BERT-based model when investigating the relation between document length and prediction performance. Additionally, as the used BERT

[4] pytorch.org

model was pre-trained on German legal documents, a domain mismatch also seems unlikely. The most likely reason for the lower performance on our dataset seems to be the significant skew in the distribution of labels. Targets with many samples are usually easier to predict, but the BERT-based model seems to need many more samples per class in many cases. Still, many samples do not necessarily result in high prediction scores, as, for example, the trade law (HGB) holds 974 samples in the train set and is never predicted for the reduced test cases. For all norms that are not predicted (241 out of the 328 reduced norms in the test set), it has the highest number of samples, but there are also norms with a lower sample count and near-perfect predictions, e.g., the law on associations (VereinsG) with 13 train samples and an F1 score of 100%. More specific laws with more minor possible applications might be easier to predict, but we did not investigate this further for the BERT-based model. Next, we will closely look at our best model based on TF/IDF and discover why it can actually outperform the other models before pointing out possible problems.

Due to the number of possible targets it is not feasible to manually identify specific shortcomings for each particular target. Nevertheless, as our best model is rather simple, it is possible to reason about its predictions and performance. In our case, y_i is the prediction value for the i-th norm $n_i \in N$, $t \in T$ are the individual terms, b_i a bias term towards predictions for a specific norm and w_t^i is the weighting of the term t for predicting the i-th norm. With σ denoting Sigmoid function used for calculating a prediction value between $[0,1]$ and $f(t)$ as the TF/IDF value our prediction function can be written down as follows:

$$y_i = \sigma(\sum_{t \in T} w_t^i f(t) + b_i) \tag{1}$$

As the TF/IDF-based model assigns a positive or negative weight w_t^i to the terms used in a court ruling (which is weighted by importance via TF/IDF) and as those individual weightings are then additively aggregated, we can identify which terms have a positive or negative impact on a prediction. When looking at the magnitude of those weights $\{w_t^i | n_i \in N\}$, we can identify that a significant majority of all terms is negatively weighted. This makes intuitively sense as for all targets, there are more negative than positive samples, and the model needs to focus more on which targets not to predict.

The resulting follow-up question is which terms are the most negatively/positively weighted, and is there an overlap between the different targets. We only want to focus on the reduced target set for this analysis, as this is not feasible otherwise. When overlapping the five positive and negative weights with the highest magnitude, we can identify in which cases a specific cue word is a positive sign for one target and a negative for another. Thus we can identify which target clusters have a thematic overlap but need some differentiation via negative cue words. We can interpret the weights like this as a big negative weighting towards a term that will only happen due to a repeated misclassification towards a different target. Moreover, we can identify these different targets by looking at enormous positive weights as they have the most decisive influence on classification.

To visualize this, we build a network with the reduced norms, where two norms A and B are connected if A has at least one term in its top-5 positive terms, which is also in the top-5 negative terms of B. This network of connected norms can be seen in Figure 2. There are smaller components of two or three norms that are connected, such as the property tax law (VStG) and the law for taxation in foreign relations (AStG). As far as

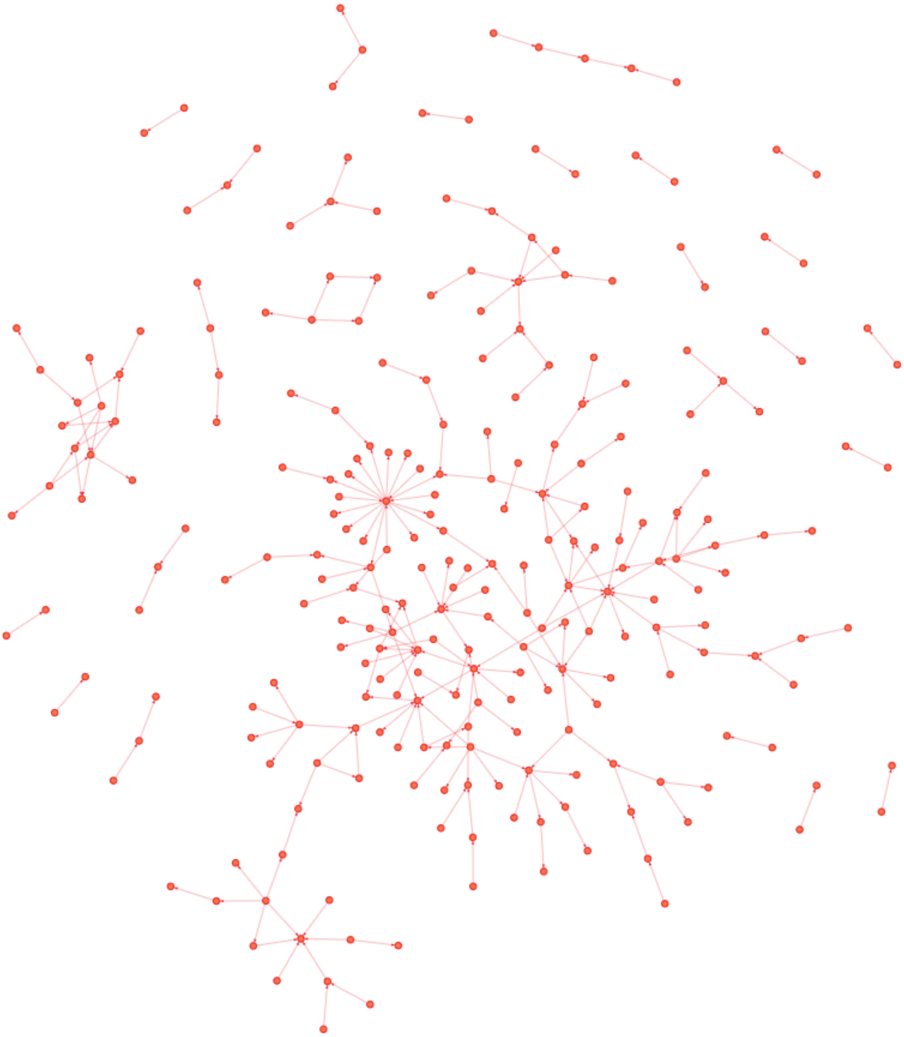


Figure 2. Connection graph of reduced norms which share at least one term in their top-5 positive and negative weights for the TF/IDF-based model. Unconnected norms were omitted.

we can tell, those connections denote exceptional cases for specific circumstances, such as, in this case, the location change. Furthermore, in those cases, one of the norms is not applicable anymore. More prominent connected components and especially star-shaped connections are interesting cases. For example, the bigger cluster on the left deals with taxation laws, with most of the norms specifically focus on energy taxes for natural gas or combined heat and power generation. In those cases, the prediction depends on precise textual details. However, by far, the most significant connected component is around the federal constitutional court law (BVerfGG), with the star in the middle of the picture. As this is only applicable at the highest instance level, our model needed to implicitly identify some information about the instance level of a court ruling to assign this law correctly. Furthermore, this becomes more evident when looking at the top-10 weights

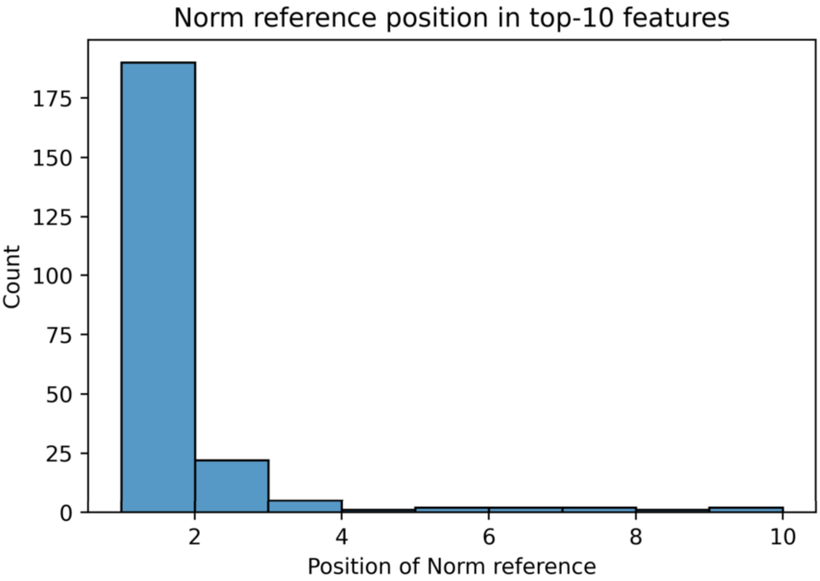


Figure 3. Position of the norm reference in the top-10 features for the reduced norms. Average position is 1.4 and only 84 norm targets do not have their reference in the top-10 positive terms.

with multiple dozen connections to this law.

We now want to look at the positively weighted terms with the highest magnitude, and surprisingly the reference to the norm itself within the text is possibly the most important term. As seen in Figure 3, it is on average the 1.4th highest weighted term, and it is for only 84 of the 666 reduced norms not in the top-10 features. This fact also explains the high performance of the model with the used norms as features.

We cannot look at each essential individual term, but we can identify if they are important in the sourcing norm or if there exists a different norm for which the top-10 positive terms are more critical/higher weighted. In order to extract the most relevant terms per reference norm A, we used TF/IDF on the source documents and then extracted the values for the top-10 positive weighted terms for that norm. After extracting the values for the same top-10 terms for every other norm, we ranked all the norms based on the sum of their TF/IDF values in decreasing order. Consequently, a norm will have a high ranking if those top-10 terms are essential in its text. We then calculated the rank for the reference norm A and could identify that in the majority of cases, the reference norm has the highest summed value as seen in Figure 4. Based on this, we can tell that the prediction is based on important terms from the source document.

Lastly, we want to look at the influence of the number of trainset samples on the test set F1 score to identify how many samples would be necessary for each target norm to get more reliable predictions. As the number of occurrences per target follows a power-law distribution, we plotted the test set F1 score based on the sorted, decreasing train set occurrences for each reduced norm in Figure 5. As seen in this figure, most reduced norms with at least 50 samples in the train set can be predicted to a certain extent, but there are also cases with much fewer samples and near-perfect predictions. We can also

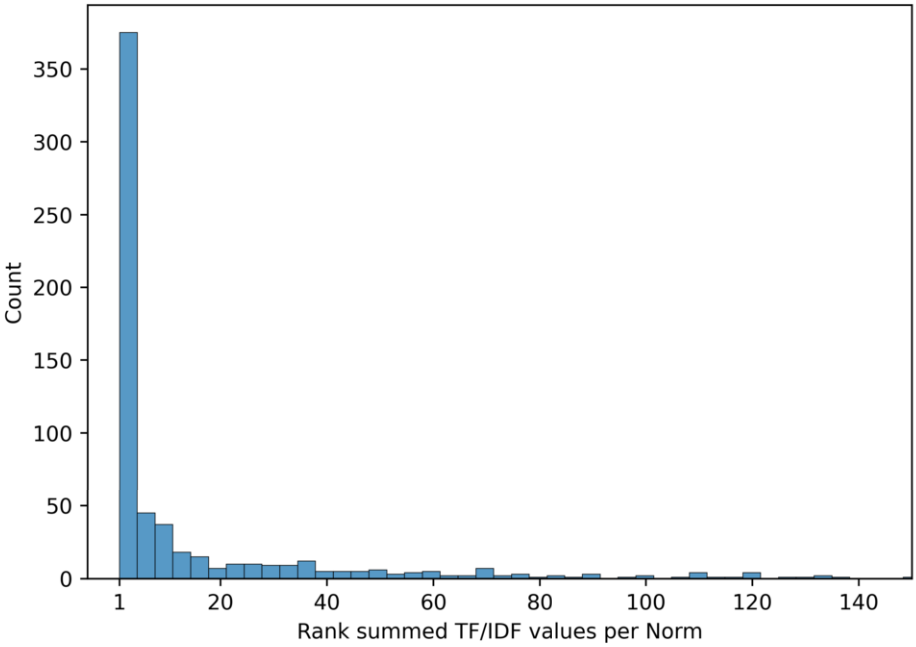


Figure 4. Ranks of the norm when extracting the TF/IDF values from the referenced norm text for the top-10 positive features. Those values are then summed and sorted in decreasing order. The rank of a norm is its position in this sorted list. Cutoff at rank 150 as the counts stay relatively similar. Maximal rank for German civil code (BGB) with rank 366.

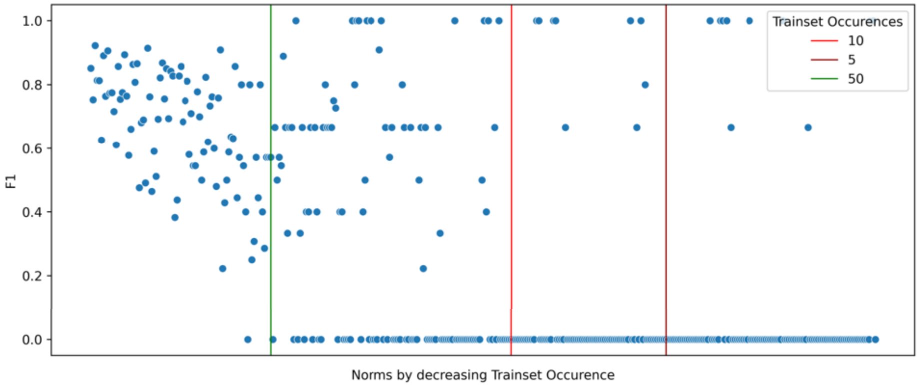


Figure 5. Number of occurrences of a reduced norm in the train set against its test set F1 score.

observe that the spread in F1 scores steadily increases with decreasing number of train samples, which was to be expected. There are again targets that are not used for any predictions, but the TF/IDF-based model uses 86 more norms in their predictions than the BERT-based model and thus at least tries to predict 53% of the possible reduced targets (173 out of 328 reduced norms in the test set). We also investigated prediction errors for the combination of norms as targets, e.g., are norms mispredicted more often when a semantically similar norm is also a target label due to a thematic overlap. We could not identify any relationship. Consequently, in our opinion, the most considerable influence on low-performance metrics for a specific (reduced) norm is the low number of samples.

When doing the same analysis with the whole target set, the results are similar, although the influence of the number of trainset samples is more extreme. As the TF/IDF model depends on some particular textual cues, we hypothesize that BERT cannot extract those specific cues.

5. Limitations and Related Work

The framework we built around our task has some shortcomings, which we want to address in the following. First, as we phrased our problem as a classification task, we cannot predict norms or reduced norms that are not in our dataset. If we want to include new norms, we would need to retrain all of our models. This is particularly problematic for norms as there are many more targets. Second, a substantial number of samples is necessary per target to get reliable predictions. As the targets follow a power-law distribution, this is not practically feasible. Third, the law is constantly changing. To be as precise as possible with our predictions, we would need to keep track of all the changes in the past and then only predict viable norms based on the date of a court ruling. Initially, we tried to avoid the first two shortcomings by posing our problem in the framework of a recommendation system with different features representations for the court rulings and norms, different extraction granularities (document level, sentence level, etc.), and different extraction methods such as approximate k-Nearest Neighbor. Thereby, we could not achieve satisfactory performance levels even when drastically increasing the number of recommendations.

Looking at related research, many papers exist that deal with multi-label text classification. As the legal domain has longer and more complex texts, we want to focus on legal text classification papers. While the list of tremendous research activities within the AI&Law community is extensive, we tried to point towards the most relevant papers for our work.

Sulea et al. [6] try to predict the decisions of the rulings of the French Supreme Court as well as its area of law, while Soh et al. [7] used classification methods on Singapore Supreme Court judgments to identify their legal areas. While much current work focuses on various classification topics such as legal document segmentation and including metadata extraction [8,9,10,11,12], Chalkidis et al. [13] are the most closely related to ours, mainly because they encountered a problem with a considerable number of possible target labels as well. They apply different Deep Learning architectures to classify European legislative documents with over 4k labels. In contrast to our work, they could achieve their highest performance with a similar BERT-based model to ours. Most recently, Huang et al. [14] provided approaches to recommend legal citations based on deep learning. They trained four different types of machine learning models. In future work, it may be worth investigating whether their approach can be applied to court rulings and norms to identify possibly relevant norms. Such identification of potential norms could then be utilized as another feature in our classification models.

6. Conclusion

This paper examined the possibility of automating the legal norm chain creation process for the German legal domain. The problem was modeled as a multi-label classification task, utilizing a linear layer for the actual classification. Four different methods were applied for the underlying featurization: (1) word2vec, (2) one-hot encoded occurrences of the individual norms, (3) TF/IDF, and (4) BERT. We could show that it is feasible up to a certain degree to extract the relevant norms from court rulings with an F1 measure of 0.77.

Nonetheless, as stated in Section 5 this research contains some limitations. Most limitations arise from the multi-label classification setup. However, we already implemented different recommendation systems in an unsupervised fashion, utilizing external sources such as vectorized representations of the actual norm content. As a result, we believe that the task must be tackled as a supervised multi-label classification task. It would be inevitable to capture even more different norms in future work by utilizing a larger corpus. In doing so, it may be beneficial to train different models for verdicts from courts of different jurisdictions.

With this work, we laid down essential groundwork in the context of extracting the most relevant norms from court rulings. We not only provided methods that can extract such norms but also performed a detailed analysis of our models. Those insights can be utilized within the research community in future work. Moreover, our algorithms can already be integrated into the existing editorial process within legal publishers in order to provide their legal authors with adequate suggestions for norms that should be included in the norm chain.

References

[1] S. A. Lastres, “Rebooting legal research in a digital age,” 2015.

[2] L. F. Peoples, “The death of the digest and the pitfalls of electronic research: what is the modern legal researcher to do,” *Law Libr. J.*, vol. 97, p. 661, 2005.

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS’13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 3111–3119.

[4] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[6] O.-M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. Dinu, and J. Genabith, “Exploring the use of text classification in the legal domain,” 10 2017.

[7] J. Soh, H. K. Lim, and I. E. Chai, “Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments,” in *Proceedings of the Natural Legal Language Processing Workshop 2019*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 67–77. [Online]. Available: https://aclanthology.org/W19-2208

[8] Q. Lu, J. G. Conrad, K. Al-Kofahi, and W. Keenan, “Legal document clustering with built-in topic segmentation,” in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 383–392.

[9] A. Lyte and K. Branting, “Document segmentation labeling techniques for court filings.” in *ASAIL@ ICAIL*, 2019.

[10] E. Loza Mencía, “Segmentation of legal documents,” in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 2009, pp. 88–97.

[11] B. Waltl, G. Bonczek, E. Scepankova, and F. Matthes, “Semantic types of legal norms in german laws: classification and analysis using local linear explanations,” *Artificial Intelligence and Law*, vol. 27, no. 1, pp. 43–71, 2019.

[12] I. Chalkidis and D. Kampas, “Deep learning in law: early adaptation and legal word embeddings trained on large corpora,” *Artificial Intelligence and Law*, vol. 27, no. 2, pp. 171–198, 2019.

[13] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, “Large-scale multi-label text classification on EU legislation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6314–6322. [Online]. Available: https://aclanthology.org/P19-1636

[14] Z. Huang, C. Low, M. Teng, H. Zhang, D. E. Ho, M. S. Krass, and M. Grabmair, “Context-aware legal citation recommendation using deep learning,” *arXiv preprint arXiv:2106.10776*, 2021.