Classification of German Court Rulings: Detecting the Area of Law

Ingo GLASER [a], and Florian MATTHES [a]

[a] *Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Germany*

Abstract. This paper investigates on the feasibility of automatically detecting the legal area of court rulings. Hereby, we establish the hypothesis that the allocation to a field of law is often ambiguous and errors occur in that process as a result. A dataset constituting over 9.000 labelled court rulings was used in order to train different machine learning (ML) classifiers. Additionally, we applied rule-based approaches utilizing domain knowledge of legal experts. Our models outperformed the rule-based approaches significantly. Hence, we could show that the performance of ML models are less prone to errors than the manual assignment of legal experts.

Keywords. legal document classification, area of law detection, semantic analysis of court rulings, natural language processing

1. Introduction

Legal rules are only partially defined in legislation. Besides, in the context of jurisdiction, there is a permanent local, case-related clarification and concretization of the terms used in legislative texts. Court rulings, therefore, represent, next to legislative texts, a second important source of conceptual knowledge for the practice of law.

Hence, court rulings play not only an important role in legal research but are also part of the daily work of legal practitioners. Various online databases exist to make judgments accessible in the digital age. These databases aim at offering state-of-the-art information retrieval features to put useful search functionalities at their disposal. Therefore, the court rulings are enriched with semantic information. That is when the work of legal authors begin.

One crucial piece of information is the area of law on which the decision is based. At first glance, it seems easy for a legal author to decide to which area of law a given verdict belongs. However, as this is only one of the tasks of a legal author when enriching court rulings with semantic information, every automation allows one to focus on other tedious and time-consuming tasks (e.g. building norm chains, writing guiding principles, etc.). Furthermore, we establish the hypothesis that the allocation to a field of law is often ambiguous, and errors occur in that process as a result. Hence, in this paper, we want to investigate the feasibility of automatically detecting the legal area of court rulings.

The remainder of the paper is structured as follows: Section 2 provides a short overview of the related work, Section 3 describes the legal areas leveraged, the experi-

Proceedings of the 2020 Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL)

mental setup along with the used dataset is discussed in Section 4, finally, the approaches and its performance is evaluated in Section 5, before Section 6 closes with a conclusion and outlook.

2. Related Work

The computer-assisted semantic analysis of court rulings is highly relevant and has attracted researchers for quite some time. However, hardly any attempt has been made in the German legal domain. Waltl et al. [1] attempted to predict the outcome of appeal decisions in Germany's tax law. They trained different machine learning classifiers based on the previous instance to determine likelihood ratios and thus predict the outcome of the appeal. In another paper from 2017, Waltl et al. [2] demonstrated the rule-based extraction of semantic information, such as the year of dispute, from court rulings in the area of tax law.

However, approaches to utilize court rulings for various analyses exist in different jurisdictions. An important contribution concerning pre-processing court rulings was made by Savelka et al. [3]. They showed that legal decisions are more challenging for existing sentence boundary detection (SBD) systems than for non-legal texts and trained conditional random fields (CRF) and outperformed state-of-the-art SBD systems when applied to adjudicatory decisions. Westerman et al. [4] have built classifiers in the form of boolean search rules on four different legal datasets, including statutory data, to provide an explainable legal classification setup. As it is a core activity in legal decision making, the identification of relevant or similar court decisions was investigated by Moodley et al. [5]. Hereby, the authors compared the results of state-of-the-art text similarity algorithms with the citation behavior in the case citation network for the Court of Justice of the European Union. The fact that labeled datasets in the legal domain are most often small, scarce, and expensive was taken up by Condevaux et al. [6] as they utilized weakly supervised one-shot classification for recurrent neural networks (RNN) to overcome the issue of data scarcity. In their work, they focused on predicting the outcome of decisions given highly ambiguous judge arguments. Slingerland et al. [7] classify Dutch civil law judgments based on whether they involve the Brussels I Regulation, including the Recast, or not.

As Brueninghaus and Ashley [8] explained that the desire of attorneys to find the most relevant cases caused the broad interest of text classification in the legal domain, many research was conducted applying legal text classification [9,10,11,12,13,14]. While there exists related work with regard to the classification of legal texts, to the best of our knowledge, there is no work concerning the detection of the area of law of German court rulings.

3. Area of Laws

The classification of legal court rulings into different areas seems obvious at first glance. However, such a classification highly depends on the specific domain.

For this work, we utilized common areas of law that were provided by a German legal publisher who uses them in practice. Their classification system consists of 92 different classes. Since this classification is very fine-grained and therefore few examples per class are available, we needed a more abstract classification. Table 1 reveals the taxonomy used in this work.

The first column depicts the 16 fields of law that are utilized in our experiments. There is a null class to intercept decisions that do not fall under any of the specific areas of law. While there may be room for arguing whether some of the classess are mapped properly, the underlying system has been proven in practice as it is employed by many legal magazines.

Furthermore, we came up with a scenario constituting only four different fields of law. The mapping between the 16 and four class setup is revealed in Table 2, and its reasoning is discussed in Section 4.1.

4. Experimental Setup

Legal research is a crucial task in the field of law. Due to the tedious and labor-intensive work, we are investigating the automatic classification of court rulings according to different areas of law. We not only utilize machine learning (ML) to classify decisions into fields of law but also mimic human behavior on that task by capturing it in a rule-based manner.

4.1. Data

For this research, we utilized a dataset provided by a German legal publisher. The dataset consists of 9,563 civil law decisions from 2010 to 2020 in XML format constituting various German court instances (e.g. Supreme Court, Regional Courts, Higher Regional Courts, Local Court) in the ordinary jurisdiction. In a first preprocessing step, the raw text of these decisions was extracted from the XML files. Since the XML already contained a segmentation into the components of a judgment, it was possible to retrieve these separately:

1. *Heading (Rubrum):* The so-called rubrum is, so to speak, the introduction of the judgment and consists of file number, header, name of the parties involved, in particular the plaintiff and defendant and any legal representatives, the court, date of the last trial, and designation of the judgment.
2. *Guiding Principle (Leitsatz):* A guiding principle is not directly part of the court ruling, but is prepared upon publication and represents a summary of the main reasons for the decision by the court.
3. *Tenor (Tenor):* The most important part of the judgment, as this is where the legal dispute is decided, so i.e. whether the defendant is sentenced to the plaintiff the sued amount or whether the claim is dismissed. The tenor is composed of three things: the actual tenor, the decision on costs, and provisional enforceability including the power to avert the judgment.

Area of Law	Granular Areas
Labor and Social Law	Labor Law (Overall); Labor Law, Social Law (Other); Company Pension Scheme; Individual Labor Law; International / European Labor Law; Collective Labor Law; Social Security Law;
Banking and Credit Security Law	Banking and Credit Security Law;
Building Law	Building Law; Brokerage and Property Development Law; Private Building Law;
Professional Law	Professional Practice; BNotO / BeurkG / DONot; Law Firm Management;
Family and Inheritance Law	Care, Accommodation; Marriage and Divorce Law; Matrimonial Property Law; Law of Succession; Family Law; International Family :aw; Childhood, Descent, Adoption; Maintenance; Supply Equalization;
Toll Law	Fee Law, Cost Law; Costs and Fees; Notary Fees;
Commercial and Corporate Law	Stock Corporation Law; Corporate Law (General); Corporate Law (Other); GmbH & Co. KG; GmbH Law; Commercial Law; International Company Law; Partnership Law; Partnership and Corporate Law; Law of Associations;
Liability and Insurance Law	Public Liability Law; Liability; Liability Law; Medical Law; Insurance Law;
Tenancy and Real Estate Law	Real Estate Law; Tenancy Law, Lease Law; Condominium Law;
Enforcement and Insolvency	Insolvency Law; Execution Law;
Motor Vehicle and Traffic Law	Motor Vehicle Law; Traffic Law;
Neighbourhood Law	Neighbourhood Law;
Procedural Law	Out-of-court Conflict Resolution; Family Procedure Law; International Procedural Law; Arbitration Proceedings; Proceedings of Voluntary Jurisdiction; Civil Procedure Law (Other); Procedural Law; Procedural Law (General); Civil Procedure Law;
Contract Law	ToS Law; IPR; Right of Purchase; Leasing; Travel Law; Contract Law / ToS Law;
Competition Law and Industrial Property Rights	Antitrust Law; Trademark Law; Patent Law; Pharmaceutical Law; Competition Law;
Other	Data Protection Law; Energy Law; European Civil Law; Advanced Training; International Business Law; Internet Law; IT Law, Media Law (Other); Public Commercial Law; Press Law; Property Law; Tax Law (General); Criminal Procedural Law (Other); Civil Law (General); Telecommunications Law; Transport Law; Conversion Right; Copyright; Constitutional Law; Public Procurement Law; Administrative Law; Civil Law (Other); Business Law (General); Commercial Law (Other);

Table 1. The used taxonomy provided by a German legal publisher

4. *Facts (Tatbestand):* In it, the facts on which the judgment was based are presented in the same way as they were presented to the court after the last oral hearing.
5. *Reasoning (Entscheidungsgründe):* The court states the reasons for its decision in the reasoning part.

The separation into these components allowed us to investigate different classification inputs. Even though the tenor is mandatory for every German court ruling (117 II Nr. 3 VwGO), 69 decisions did not include a tenor. The same applies to the reasoning part (117 II Nr. 5 VwGO). 11 court rulings that do not contain any reasoning can be explained by the fact that these are guiding principle decisions. Often the facts of a given case are summarized under the reasoning part. Therefore only 3,929 decisions explicitly state the facts. Last but not least, four decisions did miss the respective label and thus, were removed from the dataset.

As the data has been published in various legal magazines, the decisions were labeled over the course of many years according to the company guidelines of the legal publisher. The distribution of the different fields of law for the remaining 9,550 verdicts is revealed in Table 2.

4-class Setup	Area of Law	#	Relative (%)
Business Law	Banking and Credit Security Law	207	2.2
	Commercial and Corporate Law	525	5.5
	Competition Law and Ind. Prop. Rights	607	6.4
	Enforcement and Insolvency	880	9.2
	Liability and Insurance Law	1,002	10.4
		∑ 3,221	33.7
Labor Law	Labor and Social Law	450	4.7
Tenancy Law	Tenancy and Real Estate Law	926	9.7
Other	Building Law	243	2.5
	Contract Law	592	6.2
	Family and Inheritance Law	1,196	12.5
	Motor Vehicle and Traffic Law	313	3.4
	Neighbourhood Law	50	0.5
	Procedural Law	1,808	18.9
	Professional Law	172	1.8
	Toll Law	469	4.9
	Other	119	1.2
		∑ 4,962	51.9
		∑ 9,559	100

Table 2. Support of the different fields of law in the dataset

4.2. Experiment

To be able to assess our two hypotheses, our experimental setting constituted three steps:

1. *Rule-based Classification with Four Classes:* We implemented a rule-based approach and evaluated it using different parts of the dataset.

2. *Model Training with Four Classes:* We trained various classifiers on the dataset with four classes and evaluated them using 10-fold cross-validation on 20% of the data.
3. *Model Training with 16 Classes:* We trained various classifiers on the dataset with 16 classes and evaluated them through 10-fold cross-validation on 20% of the data.

4.2.1. Rule-based classification

Three documents describing the approach to manually identify the area of law for a given court ruling built the foundation for the rule-based approach. These documents were created by three legal experts as they perform this task daily. At first, we transformed the descriptions into rules employing the programming language Python. These rules consist of four distinct criteria:

1. *Specific Laws (SL):* We extracted all legal references and compared them against a list of laws that is typical for the underlying area.
2. *Specific Norms (SN):* Even one level deeper, we looked at certain norms.
3. *Typical terms (TT):* A simple lookup to check for the occurrence of specific terms.
4. *Cited Literature (CL):* Based on the extracted legal references, the occurrence of certain literature such as commentaries was counted.

Furthermore, the file number of the decision provides valuable information. Higher courts such as the German Supreme Court (BGH) utilize a unique system for the file numbers. Such a file number indicates the underlying field of law. This information was taken into account as well. The rule-based algorithm favors the file number, i.e. if the decision origins from a court with such a file number system, the classification is done purely based on that number.

In all other cases, the algorithm counts the occurrences of the specific laws, norms, terms, and literature. The weights used in the formula were provided by the legal experts. Figure 1 depicts that approach.

$$Score = SL + SN * 2 + TT * 0.5 + CL * 3$$

Figure 1. Formula for the score of the rule-based approach

We then applied different thresholds to the score to classify each decision into one field of law. Section 5.1 elaborates on these thresholds in greater detail.

4.2.2. ML-based classification

As discussed already, two different settings were applied. However, since the only difference is the underlying variation of the taxonomy, the same approaches were used for both setups. Therefore, the following steps were implemented:

Pre-Processing: We used three different pre-processing procedures: (1) The normalization (PRE) consisted of the removal of line breaks as well as duplicated whitespaces, replacing German umlauts, spelling numbers, and removing punctuation. (2) Stop word removal (SWR) was performed according to the spaCy[1] stop word list. (3) A lemmatization (Lemma) was conducted leveraging spaCy. These three procedures were incorporated into pipelines in different combinations.

Feature Extraction: Four different feature representations were used: (1) A bag-of-words approach was utilized to represent features. We used simple word count vectors as well as where indicated, additionally a term frequency-inverse document frequency (TFIDF) transformer on these vectors. Where indicated, part-of-speech (POS) tags have been created and used as well. To keep the bag-of-words approach in this case as well, each token was combined with the respective POS tag using a dash. (2) The second feature representation leveraged word embeddings. We trained word2vec models on different legal corpora as well as used pre-trained models. These models were used to calculate the mean embedding of a decision component (e.g. reasoning). (3) We also incorporated topic modeling to create features. (4) Finally, we utilized state-of-the-art deep neural representations such as BERT [15].

Training of Machine Learning Model: Six different traditional classifiers were applied to the task of predicting the legal area of court rulings. We used multinomial naive Bayes (MNB), logistic regression (LR), support vector machines (SVM), multilayer perceptrons (P), random forests (RF), and an extra tree classifier (ETC). The models were trained using 10-fold cross-validation on 80% of the dataset each iteration. Furthermore, we trained different deep neural architectures on 60% of the data. Hereby, 20% of the data was used for validating during training, while the other 20% acted as a hold-out set for the final testing.

Evaluation and Error Analysis: Weighted variants of precision, recall, and F1 was used to evaluate the performance of the trained models.

5. Evaluation & Error Analysis

5.1. Evaluating the Performance

The objective of this work was to evaluate the possibility of automatically detecting the field of law for a given legal verdict as well as to compare ML-based with rule-based approaches.

To achieve this, different classifiers were incorporated into various pipeline settings as described in Section 4.2.2 and applied to the dataset constituting four classes. While this resulted in over 50 different models, we also utilized state-of-the-art deep neural networks utilizing contextual embeddings such as BERT [15]. In Table 3 we only report on the best performing classifier for space reasons.

That is a SVM applied on a pipeline performing our pre-processing procedure utilizing TFIDF on the lemmatized input. As input, only the reasoning part of the court rulings was used as the other components resulted in inferior performances. The resulting model performed with an F_1 of 0.87. In contrast, the initial rule-based approach only achieved

[1] https://spacy.io/usage/linguistic-features

Method	Class	Precision	Recall	F_1
RB: Biggest Score	Business Law	0.21	0.97	0.34
	Labor Law	0.68	0.61	0.64
	Tenancy Law	0.66	0.63	0.64
	Other	0.86	0.10	0.21
	Weighted Avg.	0.72	0.25	0.15
RB: Minimum of 30	Business Law	0.31	0.50	0.38
	Labor Law	0.68	0.61	0.64
	Tenancy Law	0.66	0.63	0.64
	Other	0.81	0.69	0.75
	Weighted Avg.	0.70	0.65	0.67
SVM TFIDF (PRE + Lemma)	Business Law	0.71	0.75	0.73
	Labor Law	0.78	0.93	0.85
	Tenancy Law	0.78	0.74	0.76
	Other	0.92	0.91	0.92
	Weighted Avg.	0.88	0.87	0.87

Table 3. Selected performances of the approaches applied on the four class dataset

an F_1 of 0.15. The bad result can be attributed to the low recall, particularly for the null class. As a result, we introduced a threshold for the score, i.e. the score needs to meet a defined value to be eligible for classification. If this condition is not met, the decision will be classified into the class *Other*. As one can see in Table 3 due to that threshold the recall was increased while keeping the precision almost consistent. We achieved the best result with a threshold value of 30 as a bigger threshold worsened the performance.

The results already suggest that both our hypothesis can be proved: (1) It is possible to automatically detect the field of law for a given court ruling with good performance (88%), and (2) ML-based approaches outperform rules mimicking human behavior.

Nonetheless, a setup consisting of only four classes is not usable in practice. For that reason, we selected the best performing setup and trained it on the dataset with 16 classes as well. Table 4 reveals the resulting performance.

As the overall performance decreases significantly (F_1 0.75), the approach does not seem to generalize well at first glance. However, it can be seen that the performance only drops in underrepresented classes such as *Professional Law* or *Neighbourhood Law*. Classes with high support, e.g. *Family and Inheritance Law* or *Enforcement and Insolvency* perform almost on the same level as in the four-class setting. This suggests that adding sufficient data leads to results of around 90% as well.

5.2. Error Analysis

The comparison of the results in Table 3 provides evidence in our initial hypotheses, which stated that machines utilizing ML are more capable of extracting the area of law from court rulings. To be able to better understand the differences between our rules and the ML models, the best configuration (SVM utilizing TFIDF with our pre-processing and lemmatization) was examined in greater detail.

Method	Class	Precision	Recall	F_1	Support
SVM (PRE + Lemma)	Labor and Social Law	0.93	0.89	0.91	45
	Banking and Credit Security Law	0.65	0.62	0.63	21
	Building Law	0.83	0.62	0.71	24
	Professional Law	0.80	0.24	0.36	17
	Family and Inheritance Law	0.79	0.87	0.82	119
	Toll Law	0.60	0.55	0.58	47
	Commercial and Corporate Law	0.76	0.79	0.78	53
	Liability and Insurance Law	0.76	0.81	0.78	100
	Tenancy and Real Estate Law	0.79	0.82	0.80	92
	Enforcement and Insolvency	0.88	0.88	0.88	88
	Motor Vehicle and Traffic Law	0.82	0.74	0.78	31
	Neighbourhood Law	0.50	0.20	0.29	5
	Procedural Law	0.68	0.75	0.71	181
	Contract Law	0.75	0.64	0.69	59
	Competiton Law and Ind. Prop. Rights	0.75	0.85	0.80	61
	Other	0.1	0.1	0.1	12
	Weighted Avg.	0.75	0.76	0.75	955

Table 4. Performance of the best performing model applied on the 16 class dataset

We looked into the existing model and inspected the coefficients of each feature. The most important features for the class *Labor Law* are (1) *landesarbeitsgericht*, (2) *arbeitnehmer*, (3) *arbeitgeber*, and (4) *arbeitsverhältnis*. When looking at the rules for the classification of a decision into *Labor Law*, these terms occur in our *TT*. However, these typical terms only play a small role when assigning the classes through the rules. In general could be observed that the rules focus heavily on references to specific laws and norms, while all of our ML models mostly pay attention to specific terms. Even greater evidence is provided when revealing the most important features for the classes *Tenancy Law* and *Business Law*. While the former uses words representing typical terms, the latter focuses on laws as well (e.g. *hgb*, *aktg* or *inso*) resulting in a worse performance.

6. Conclusion & Outlook

In this paper, we investigated the possibility to automatically detect the area of law for a given legal judgment. Therefore, various classifiers were trained on a dataset constituting 9,559 German court rulings. While some simple linear models were created, also advanced techniques such as topic modeling were incorporated.

We could make two contributions: (1) Highlighting the ability to automatically detect the area of law for German legal court rulings with high precision, and (2) show that ML-based approaches outperform a human inspired rule-based baseline.

Nonetheless, this research includes some limitations. The verdict components were of different lengths. Furthermore, the class distribution varied quite a lot. As a consequence, future research needs to define an even more suitable setting in terms of data distribution and size to provide more evidence on the capability of ML models. Yet, this work builds a solid base for future research in this area.

While we could not achieve great results utilizing state-of-the-art contextual embeddings, usually such approaches are superior. As a result, it may be worth another attempt at selecting different deep neural architectures relying on contextual embeddings.

References

[1] B. Waltl, G. Bonczek, E. Scepankova, J. Landthaler, and F. Matthes, "Predicting the outcome of appeal decisions in germany's tax law," in *Electronic Participation*, P. Parycek, Y. Charalabidis, A. V. Chugunov, P. Panagiotopoulos, T. A. Pardo, Ø. Sæbø, and E. Tambouris, Eds. Cham: Springer International Publishing, 2017, pp. 89–99.

[2] B. Waltl, J. Landthaler, E. Scepankova, F. Matthes, T. Geiger, C. Stocker, and C. Schneider, "Automated extraction of semantic information from german legal documents," in *IRIS: Internationales Rechtsinformatik Symposium*, 2017.

[3] J. Savelka, V. R. Walker, M. Grabmair, and K. D. Ashley, "Sentence boundary detection in adjudicatory decisions in the united states," *TRAITEMENT AUTOMATIQUE DES LANGUES*, vol. 58, no. 2, pp. 21–45, 2017.

[4] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, and K. Benyekhlef, "Computer-assisted creation of boolean search rules for text classification in the legal domain." in *JURIX*, 2019, pp. 123–132.

[5] K. Moodley, P. V. H. Serrano, G. van Dijck, and M. Dumontier, "Similarity and relevance of court decisions: A computational study on cjeu cases," in *JURIX*, 2019, pp. 63–72.

[6] C. Condevaux, S. Harispe, S. Mussard, and G. Zambrano, "Weakly supervised one-shot classification using recurrent neural networks with attention: Application to claim acceptance detection." in *JURIX*, 2019, pp. 23–32.

[7] R. Slingerland, A. Boer, and R. Winkels, "Analysing the impact of legal change through case classification." in *JURIX*, 2018, pp. 121–130.

[8] S. Brüninghaus and K. D. Ashley, "Toward adding knowledge to learning algorithms for indexing legal cases," in *Proceedings of the 7th international conference on Artificial intelligence and law*, 1999, pp. 9–17.

[9] E. de Maat, K. Krabben, and R. Winkels, "Machine learning versus knowledge based classification of legal texts." in *JURIX*, 2010, pp. 87–96.

[10] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria, "Automatic semantics extraction in law documents," in *Proceedings of the 10th international conference on Artificial intelligence and law*, 2005, pp. 133–140.

[11] J. O. Neill, P. Buitelaar, C. Robin, and L. O. Brien, "Classifying sentential modality in legal language: a use case in financial regulations, acts and directives," in *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, 2017, pp. 159–168.

[12] I. Glaser, E. Scepankova, and F. Matthes, "Classifying semantic types of legal sentences: Portability of machine learning models." in *JURIX*, 2018, pp. 61–70.

[13] M. H. Falakmasir and K. D. Ashley, "Utilizing vector space models for identifying legal factors from text." in *JURIX*, 2017, pp. 183–192.

[14] B. Waltl, J. Muhr, I. Glaser, G. Bonczek, E. Scepankova, and F. Matthes, "Classifying legal norms with active machine learning." in *JURIX*, 2017, pp. 11–20.

[15] 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.