SUPPORTING THE LEGAL REASONING PROCESS BY CLASSIFICATION OF JUDGMENTS APPLYING ACTIVE MACHINE LEARNING

Ingo Glaser / Jörg Landthaler / Florian Matthes

Research Associate, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
ingo.glaser@tum.de, http://wwwmatthes.in.tum.de

Research Associate, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
b.waltl@tum.de, http://wwwmatthes.in.tum.de

Professor, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems
Boltzmannstraße 3, 85748 Garching bei München, DE
florian.matthes@in.tum.de, http://wwwmatthes.in.tum.de

Keywords: ***Legal Text Analysis, Legal Sentence Classification, Natural Language Processing, Legal Reasoning, Active Machine Learning***

Abstract: *The digitalization of information is transforming the way we live and creating many new business models. Digitalization is also taking place in the legal domain. Legal documents, such as contracts and general terms and conditions, are produced thousands of times a day due to numerous online contract generators, e-commerce platforms, banks and insurance companies. As a result, computer-aided legal reasoning has become an attractive research area. The purpose of this research is to investigate the applicability of active machine learning and binary legal text classification in order to detect sentences, providing a statement about the ineffectiveness of a clause.*

1. Introduction

The way we live has been changing due to the digitization of information, as well as it has been creating many new business models. Autonomous cars, internet of things, social media and artificial intelligence are just examples for a few trending technologies that make heavily use of digitally available data. In 2016, 16.1 Zettabyte (ZB) of data were generated worldwide. According to estimates, 80% of the newly generated data is unstructured [Raghavanetal et al. 2004].[1] By the year 2025, the amount of data generated is expected to rise up to 163 ZB [Reinsel et al. 2017].

Due to this increase of available unstructured data, and the enhanced capabilities of algorithms and computing power, the demand for automated data processing, e.g. text classification, pattern finding and knowledge extraction, is increasing and has become an important area for research [Khan et al. 2010]. One measure of progress in machine learning (ML) is the significant amount of existing real-world applications, like speech recognition, computer vision, robot control and accelerating empirical sciences [Mitchell 2006]. Past research has shown the successful application of various ML classification algorithms on text-based data.

[1] See https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know (all websites last accessed on 4 January 2019).

Digitization is also taking place in the legal domain. During the last legislative period (2013-2017) of the German parliament more than 550 laws were updated or created.[2] Most of the laws are available online.[3] Every year, more than 6.000 judgments are adjudicated at the German Federal Supreme Court (BGH).[4] Over 40.000 of these decisions can be accessed through an official database, that exists since 2016.[5] Other legal documents, such as contracts and general terms and conditions, are produced thousands of times a day due to numerous online contract generators and e-commerce platforms but also due to the transformation of banks and insurance companies. As a result of this information inflation, the legal domain faces new challenges, especially for judges and lawyers [Paul/Baron 2006]. This information inflation makes automatic text classification of legal texts through ML an attractive and promising research topic [Waltl et al. 2017a].

On the other side, the work of legal practitioners is time consuming and knowledge intensive. A heavy amount of a lawyer's workload focuses on legal research with regard to laws, precedent cases, and facts. Often, hereby the aim is to seek evidence to solve a present issue in favor of the client. In that process, software could help by automatic extraction of relevant predicates. Another example is the legal reasoning process of a judge. A judge strives for consistency and further development of law, while attempting to regulate social relations by applying the legislature's intention. Algorithms could help to automatically detect relevant predicates. On the other side, the society is facing contractual regulations quite often. However, unlawful clauses are frequently observed. Therefore, this paper aims at supporting the legal reasoning process, by identifying predicates within verdicts with regard to the ineffectiveness of clauses, utilizing active machine learning (AML) methods.

The remainder of this paper is structured as follows: Section 2 provides a short overview of the related work, Section 3 describes relevant legal concepts. The experimental setup along with the used dataset is discussed in Section 4, finally the configurations and its performance are evaluated in Section 5, before Section 6 closes with a conclusion and outlook.

2. Related Work

The semantic analysis of documents, by means of natural language processing (NLP), is very attractive for the legal domain. Many reasonable use cases exist, where automatic or computer-aided text processing is able to increase productivity. Ashley provided a comprehensive overview most recently [Ashley 2017].

Various research in the area of text classification within legal documents has been published recently. Particularly for the German legal domain, Waltl et al. [Waltl et al. 2017c], [Waltl et al. 2018] and Glaser et al. [Glaser et al. 2018a], [Glaser et al. 2018b] focused on the information extraction of statutory text as well as contractual text. Walter focused in his dissertation on the extraction of legal definitions within the federal constitutional court (Bundesverwaltungsgericht; BVerfG) [Walter 2009]. The linguistic variety was modelled in detail, also considering the difficult problem of negations within legal definitions and linguistic patterns indicating a definition. His linguistic work sets a base line for further research in this direction, even though the results and the technology used, have not been convincing. Furthermore, lots of research within different legislations has been made, such as the famous work by Maat et al. in 2010 [Maat et al. 2010].

Research with regard to text analytics on verdicts has been performed quite intensively as well. The extraction of arguments from legal cases was the focus of Wyner et al. [Wyner et al. 2010]. Hereby, a context-free grammar was developed, that allowed the expressions of the rules to identify those expressions. In their work, they differentiated between four classes of sentences in the context of arguments, namely premises,

[2] See http://www.bundestag.de/blob/194870/7c8a01e16c98fc9c32ddb203d7bd88e0/gesetzgebung_wp18-data.pdf.

[3] https://www.gesetze-im-internet.de/.

[4] See http://www.bundesgerichtshof.de/SharedDocs/Downloads/DE/Service/StatistikZivil/jahresstatistkZivilsenate2017.pdf?__blob=publicationFile.

[5] See https://www.bmjv.de/SharedDocs/Pressemitteilungen/DE/2016/01272016_Webservice_www_rechtsprechung_im_Internet_de_geht_online.html.

conclusions, non-argumentative information, and final decisions. The performance of their rules was diverse, revealing a precision of 89% and a recall of 80% for non-argumentative information, while the identification of premises performed worse with a precision of 59% and a recall of 70%.

Grabmair et al. created a powerful framework based on Apache UIMA in order to annotate and classify legal texts based on semantic and linguistic features [Grabmair et al. 2015]. Their work was based on a rule-based text annotation technology, called Apache Ruta, allowing a more expressive specification of linguistic patterns than simpler approaches, such as regular expressions. They extracted semantic information out of decisions regarding vaccine injuries. In that process, 49 rules were derived in order to annotate on a sentence level, as well as sub-sentence level.

With regard to German judgments, to the best of our knowledge, only two works of Waltl exist. In their first paper, they extracted semantic information, such as the year of dispute, legal definitions, and other legal concepts by means of a rule-based approach, leveraging Apache Ruta [Waltl et al. 2017]. Afterwards, they also tried to predict the outcome of appeal decisions within the German tax law, utilizing various ML models [Waltl et al. 2017b].

3. Legal Knowledge Base

This section aims to briefly introduce relevant terms and concepts within the scope of this work. First, civil cases at the highest German court, the Federal Court of Justice (Bundesgerichtshof; BGH), are explained. Secondly, the overall structure of such verdicts is discussed.

3.1. Civil Cases at the Federal Court of Justice of Germany

The BGH is a court of appeal, which means that judgments are exclusively handed to it by inferior courts for reviewing for errors of law. The remedy of appeal on points of law is only available against final judgments adopted by regional and higher regional courts acting as appellate courts. Consequently, the BGH does not perform an own fact-funding or evidence-taking. After an appeal was considered as admissible by the panel, an oral-hearing is held resulting in a written judgment. If an appeal is seen as inadmissible, it will be dismissed by way of a court order [Bundesgerichtshof 2014].

The BGH has twelve civil panels that are traditionally highly specialized for specific domains of law. In the context of this research, we only consider judgments of the eighth civil panel, who is specialized in law on the sale of goods, landlord and tenancy law.

3.2. Structure of a Civil Law Judgment

The court procedure in civil proceedings is mostly regulated by the civil procedure code (Zivilprozessordnung; ZPO), as is the general structure of a court decision in civil matters. A civil judgment is regularly divided into six parts, which are shown in Table 3.1. The remainder of this Section describes each part briefly:

(1) Recital of parties (Rubrum) The so-called recital of parties states in addition to the parties and their address, the type of judgment, the address of the court and the case reference. The case reference consists of the initials of the court, the elaborating panel of the court, a register reference and an ongoing case number succeeded by the year of receipt.

(2) Tenor The tenor forms the essence of every judgment and states the legal consequence ordered by the court, e.g., to pay the amount claimed.

Number	English	German
(1)	Recital of parties; Introduction	Rubrum
(2)	Tenor	Tenor
(3)	Summary of circumstances	Tatbestand
(4)	Opinion of the court	Entscheidungsgründe
(5)	Instruction on the right of appeal	Rechtsmittelbelehrung
(6)	Signatures of the judges	Unterschriften der Richter

Table 3.1: General structure of a German civil law judgment (own illustration based on [Hofmann 2018])

(3) Summary of circumstances (Tatbestand) The summary of circumstances reflects the essential facts after the final hearing, underlying the case, from the court's point of view. These facts are also the foundation for the final decision.

(4) Opinion of the court (Entscheidungsgründe) In addition to the opinion of the BGH, the reasoning of the lower court is also included. The argumentation of the lower court is written in indirect speech to distinguish it from the opinion of the BGH. The reasoning is written in the so-called judgment style, which begins with the result, followed by a gradual justification.

*(5) Instruction on the right of appeal (Rechtsmittelbelehrung)*According to § 232 ZPO, all civil court decisions have to include an instruction on the right of appeal, except if a legal representation is required.

*(6) Signatures of the judges (Unterschriften der Richter)*This is a separate part of the verdict, including a signature of each representative judge.

4. Experiment

The experimental setup of this research is explained in the present section. After describing the data used for the experiment, along with some pre-processing steps, the AML classification setup is described.

4.1. Data

A corpus constituting more than 800 judgments of the 8th civil senate from the BGH was used for this work. The focus of this senate is the sales of goods law as well as the tenancy law. The judgments were imported from Rechtsprechung-im-Internet[6]. The sentences resulting from this process were manually annotated by two human legal experts to serve as the training set for the AML classifier. Only sentences that are located in the tenor or the reasoning are considered (see Section 3.2), as these are the only parts of a judgment, where a statement about the ineffectiveness of contractual clauses is made. Sentences have been annotated true whenever the sentence establishes the connection between the contract clause and the legal reason of the ineffectiveness. In order to do so, for each sentence a 3-phase annotation procedure was conducted. First, it was determined, whether a contract clause was mentioned within the sentence. Afterwards, a legal justification was required in order to keep the sentence as a potential candidate for a *true* label. Eventually, the appearance of an unlawfulness clause was defined. Just if all three requirements were fulfilled, the sentence was labelled as *true*. In all other cases, the sentence was labelled as *false*. The experiment was carried out on the basis of 3.135 randomly selected sentences, of which 71 (2.26%) sentences were annotated as true. The handling of this data imbalance is discussed in the next section.

4.2. Classification

The classification setup in this work leverages an AML setting. AML is an adapted form of semi-supervised machine learning, in which the training is done in so-called learning rounds. A pre-defined number of instances

[6] https://www.rechtsprechung-im-internet.de.

from the training set are manually labelled within each round. For the selection of these instances, different query strategies exist. A more detailed description of such an AML setting, particular for the legal domain, is discussed in the work of Waltl et al. [Waltl 2017c]. The major difference is the binary classification setup, in comparison to the multiclass setup from Waltl et al. An instance is either *true* (is a predicate with regard to the ineffectiveness of a clause) or *false*. The performance of the different parameter combinations used are discussed in Section 5.1. Table 4.1 reveals some example configurations along with the chosen parameters in order to foster mutual understanding.

Name	Seed Set Size	Batch Size	Learning Rounds	Weight Factor
SS_100_QS_5_W_20	100	5	482	400
SS_100_QS_5_W_10	100	5	482	200
SS_120_QS_5_W_8	120	5	478	160
SS_60_QS_5_W_8	60	5	490	160
SS_100_QS_20_W_8	100	20	121	160

Table 4.1: Excerpt of used AML configurations

The naming of configurations adheres to the following conversion: *SS_x_QS_y_W_z*. The seed set size is specified by *x*. The seed set is the initial training set, which is required for the first training round of the classifier. The success of the AML algorithm depends heavily on the quality of the seed set. Therefore, the selection of the initial training instances is crucial. The next variable, *y*, denotes the batch size, which defines the number of instances that are queried each learning round. The standard procedure is to query one instance at a time. For knowledge-intensive classification tasks which occur for example in the legal domain, the time required to generate a model using a serial query approach is expensive. Sometimes various human annotators want to train the model at the same time. In both cases a serial query approach is unpractical. Addressing this problem, querying multiple instances at once is known as the batch mode. The primary challenge in using batch mode is finding the best Q instances. The weight factor is revealed by *y*. In order to retrieve the actual factor though, it is necessary to divide *y* by 0.05 (i.e. 20 / 0.05 = 400, as shown in the first row of Table 4.1). This weighting is necessary in order to counter the mismatch between true and false instances within the dataset, as mentioned in the previous section. Hence, the *true* instances were weighted more heavily during the learning process by utilizing the weight factor only on such instances. Additionally, Table 4.1 includes a metric called learning rounds, which indicates the number of iterations within the specific AML setting. It can be calculated by the following formula: (*Number of instances * 0.8 – seed set) / batch size*. The factor *0.8* origins from the 5-fold cross-validation applied in this work.

5. Evaluation

This section deals with the evaluation, separated into two parts. Firstly, the performance of different configurations with regard to various parameters were analyzed. Secondly, inferences concerning the support of the legal reasoning process were drawn.

5.1. Performance of different configurations

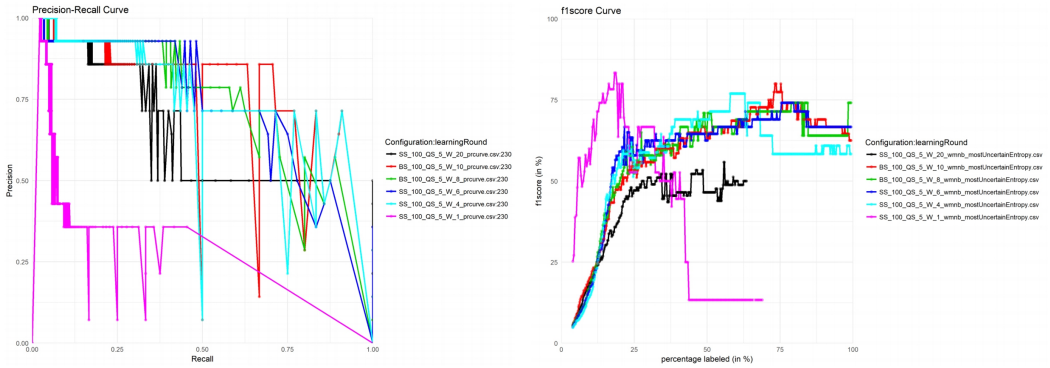

Figure 5.1: Influence of the weight

The evaluation was conducted using a 5-fold cross-validation. First of all, the selection of a suitable weight factor was crucial, as the imbalance between the two classes rises problems. Figure 5.1 reveals a precision-recall curve, as well as the respective F_1 measures utilizing different weight factors. Six different weight factors were used, by keeping a seed set size of 100, and a batch size of 5. The purple curve (*SS_100_QS_5_W_1*) does not rely on any weight factor and hence treats true and false labels with the same importance. This results in a bad curve progression, as a curve within the bottom left is rather poor, while a curve within the top right factorindicates a good performance. When looking at both curves in Figure 5.1, it becomes clear, that a very high weight factor is not helpful either. A weight factor of 160 resulted in the best results.

Based on this weight factor, the remaining experiments has been conducted. The next parameter under examination is the seed set size. Figure 5.2 depicts the performance reached by different configurations, by means of random sampling. Apparently, a bigger seed set creates better results. This seems obvious, as more training data should cause better classification results. However, Waltl et al. already showed, that AML can be superior to traditional ML approaches [Waltl et al. 2017c]. This implies after reaching a certain seed set size, adding more instances does not necessarily increase the performance. This can be seen in Figure 5.2, as a seed set size of 100 results in the best performance. Furthermore, the curves depict the so-called missed class effect.

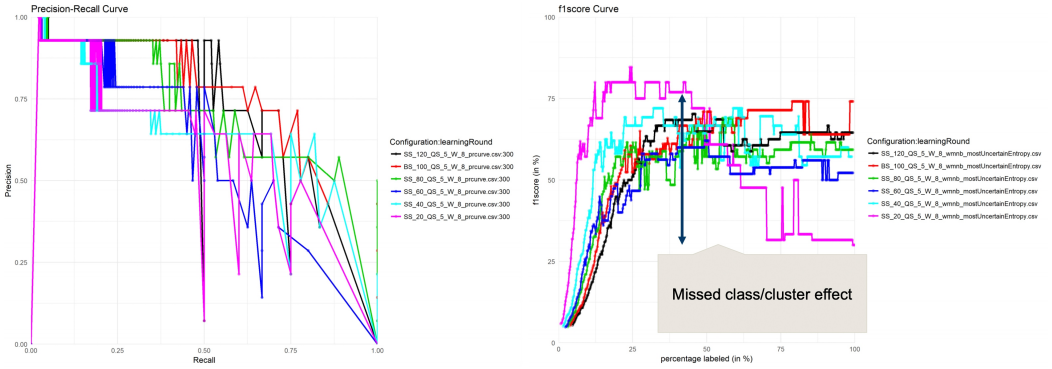

Figure 5.2: Influence of the seed set size

A small random sampled seed set is not representative in most cases and can cause a biased classification model. As a result, a sufficient exploration phase during the seed set generation is crucial.

Last but not least, the batch size was investigated. Figure 5.3 shows a heat map revealing different configurations in terms of seed set size and batch size. Again, the seed set size of 100 yields the best results. Accounting the batch size, the greatest F_1 was achieved, by using a query size of 5. As a conclusion, it can be said, that the *SS_100_QS_5_W_8* achieves the best classification results, by reaching a F_1 measure of 0.74. This configuration performs with a precision of 0.77 and a slightly worse recall of 0.71.

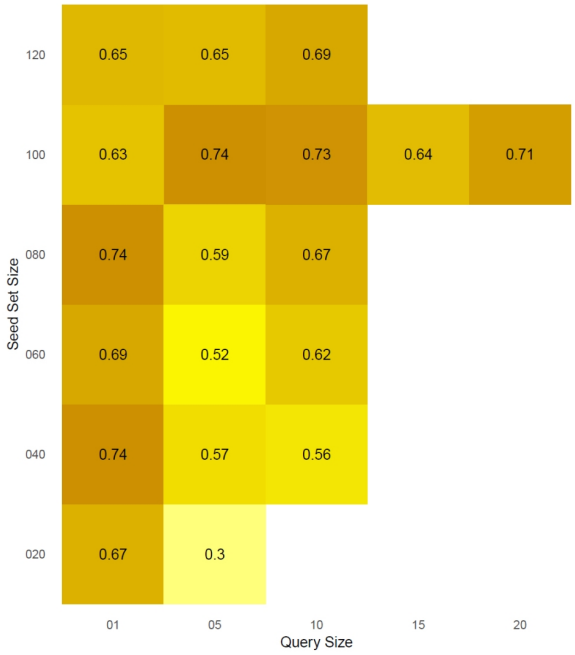

Figure 5.3: Heat map in order to measure the influence of different batch sizes

5.2. Impact for supporting the legal reasoning process

Although the empirical results collected in this study are not sufficient to properly support the legal reasoning process, existing literature has revealed some possibilities. Furthermore, this work also provides a solid baseline for further development.

Since the common law system mainly uses precedents for the legal reasoning process, lawyers often have to carry out extensive research. Today, lawyers often use online databases equipped with simple text retrieval techniques for this research. By classifying the legal reason of the ineffectiveness of contractual clause in a judgment, more far-reaching methods can be used to recognize semantic similarities. The way in which common law lawyers work is becoming more and more relevant in the European legal area as well. Parts of German law today are already heavily influenced by case law, such as tenancy law, where many rules were created by the BGH.

6. Conclusion & Outlook

In this work, only one possible use case was described, on how legal reasoning can be supported by binary text classification. For this purpose, only a limited number of approaches to the extraction of features in the context of the legal domain were utilized. This may have caused a limited performance on the performed experiments. As a result, future research should investigate different feature representations as well as data pre-processing steps.

Nonetheless, the conducted classification experiment showed that binary text classification on unbalanced classes is vulnerable for a low-quality seed set. This was largely caused due to the low quality of the data set. On the one hand, the data set was unbalanced on the other hand, the annotations made were possibly contradictory for the classifier. Section 4 described the annotation process concise, while pointing out, that it

required all three parts to be included in a sentence in order to be eligible for a true instance. Of course, this is critical, as there may be instances, where one part is mentioned in successive sentences and thus causing partially false positives.

In addition to contracts, the eighth Civil Senate of the BGH also decides on the invalidity of other declarations of intent, such as Rental contract terminations and sales contract withdrawals and revocations. One possible way to improve the use case shown could be the separation of the classification into two parts. A first classifier based on a ML or a rule-based approach would decide if the sentence has anything to do with the effectiveness of contract clauses. A second ML-based classifier would then perform the final classification task on the resulting record.

Last but not least, further steps towards a more balanced dataset are required to create even more convincing results.

7. Bibliography

Ashley, Kevin, Artificial intelligence and legal analytics: new tools for law practice in the digital age, Cambridge University Press, 2017.

Bundesgerichtshof, Der Bundesgerichtshof; the federal court of justice, Available at http://www.bundesgerichtshof.de/SharedDocs/Downloads/EN/BGH/brochure. pdf?__blob=publicationFile.

Glaser, Ingo/Waltl, Bernhard/Matthes, Florian, Named entity recognition, extraction, and linking in German legal contracts, in: IRIS: Internationales Rechtsinformatik Symposium, Salzburg, Austria, 2018.

Glaser, Ingo/Scepankova, Elena/Matthes, Florian, Classifying semantic types of legal sentences: portability of machine learning models, in: JURIX: Legal Knowledge and Information Systems, pp. 61–70, The Netherlands, 2018.

Grabmair, Matthias/Ashley, Kevin D./Chen, Ran/Sureshkumar, Preethi/Wang, Chen/Nyberg, Eric/Walker, Vern R., Introducing LUIMA: An experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools, ICAIL Proceedings, 2015.

Hofmann, Frank, Aufbau des Urteils in Zivilsachen, Repetitorium Hofmann, Freiburg, 2018.

Khan, Atif/Baharudin, Baharum/Lee, Lam Hong/Khan, Khairullah, A review of machine learning algorithms for text documents classification, Journal of advances in information technology, 1(1):4–20, 2010.

De Maat, Emile/Krabben, Kai/Winkels, Radboud, Machine learning versus knowledge based classifciation of legal texts, In JURIX: Legal Knowledge and Information Systems, pp. 87–96, 2010.

Mitchel, Tom Michael, The discipline of machine learning, vol 9, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department, Pittsburgh, 2006.

Paul, George L./Baron, Jason R., Information inflation: Can the legal system adapt, Rich, JL & Tech, 13, 1, 2006.

Raghavan, Prabhakar/Amer-Yahia, Sihem/Gravano, Luis, Structure in text: Extraction and exploitation, In Proceeding of the 7th international Workshop on the Web and Databases (WebDB), ACM SIGMOD/PODS, 2004.

Reinsel, David/Gantz, John/Rydning, John, Data age 2025: The evolution of data to life-critical, Don't Focus on Big Data, 2017.

Walter, Stephan, Definition extraction from court decisions using computational linguistic technology, Formal Linguistics and Law, vol. 212, 2009.

Waltl, Bernhard/Landthaler, Jörg/Scepankova, Elena/Matthes, Florian/Geiger, Thomas/Stocker, Christian/Schneider, Christian, Automated extraction of semantic information from German legal documents, in: IRIS: Internationales Rechtsinformatik Symposium, Salzburg, Austria, 2017.

Waltl, Bernhard/Bonczek, Georg/ Scepankova, Elena/Landthaler, Jörg/Matthes, Florian, Predicting the outcome of appeal decisions in Germany's tax law, International Federation for Information Processing (IFIP): Policy Modeling and Policy Informatics, St. Petersburg, Russia, 2017.

Waltl, Bernhard/Muhr, Johannes/Glaser, Ingo/Bonczek, Georg/Scepankova, Elena/Matthes, Florian, Classifying legal norms with active machine learning, in: JURIX: Legal Knowledge and Information Systems, pp. 11–20, Luxembourg, 2017.

Waltl, Bernhard/Bonczek, Georg/Scepankova, Elena/Matthes, Florian, Semantic types of legal norms in German laws: classification and analysis using local linear explanations, Artificial Intelligence and Law, 2018.

Wyner, Adam/Mochales-Palau, Raquel/Moens, Marie-Francine/Milward, David, Approaches to text mining arguments from legal cases, Semantic processing of legal texts, 2010.