Artificial Intelligence and Law
https://doi.org/10.1007/s10506-018-9228-y

Semantic types of legal norms in German laws: classification and analysis using local linear explanations

Bernhard Waltl[1] · **Georg Bonczek**[1] · **Elena Scepankova**[1] · **Florian Matthes**[1]

Abstract
This paper describes the automated classification of legal norms in German statutes with regard to their semantic type. We propose a semantic type taxonomy for norms in the German civil law domain consisting of nine different types focusing on functional aspects, such as Duties, Prohibitions, Permissions, etc. We performed four iterations in classifying legal norms with a rule-based approach using a manually labeled dataset, i.e., tenancy law, of the German Civil Code (n = 601). During this experiment the F_1 score continuously improved from 0.52 to 0.78. In contrast, a machine learning based approach for the classification was implemented. A performance of $F_1 = 0.83$ was reached. Traditionally, machine learning classifiers lack of transparency with regard to their decisions. We extended our approach using so-called local linear approximations, which is a novel technique to analyze and inspect a trained classifier's behavior. We can show that there are significant similarities of manually crafted knowledge, i.e., rules and pattern definitions, and the trained decision structures of machine learning approaches.

Keywords Natural language processing · Classifying legal norms · Rule-based information extraction · Supervised machine learning · Explainable machine learning · Local interpretable models

✉ Bernhard Waltl
b.waltl@tum.de

Georg Bonczek
georg.bonczek@tum.de

Elena Scepankova
elena.scepankova@tum.de

Florian Matthes
matthes@tum.de

[1] Software Engineering for Business Information Systems, Department of Informatics, Technical University of Munich, Boltzmannstraße 3, 85748 Garching bei München, Germany

Published online: 20 July 2018

1 Introduction

The computer supported analysis of legal documents is highly attractive in the legal domain (Ashley 2017). Technological advancements and increasing economical pressure on law firms facilitate the usage of digital technologies during the review of legal documents (Susskind 2013; Veith et al. 2016). There is strong evidence that common tasks such as legal research, technology assisted review (TAR), eDiscovery, forensics tasks and due diligence, which are manually performed by law practitioners and legal scientists can, at least partially, be automated by computer systems.

Within this work we describe the results from a set of experiments that have been conducted and whose objective was to semantically analyze German laws. We propose a taxonomy of semantic types for legal norms, that can be applied for statutory texts in civil law jurisdictions. It combines legal theory with empirical observations, which is underrepresented in German legal sciences. The classification of the semantics of legal norms is a challenge and foundation for many legal data science tasks and has not been addressed by German legal informatics domain yet. Recent approaches on detecting conflicts of norms within contracts by Paulo Aires et al. (2017) also dealt with the challenge to determine the semantic types of legal norms.

In order to extract the semantics of an unstructured, i.e., textual, legal document, different technologies could be used, each of which has pros and cons. Within this article we describe the results from two approaches of classifying legal norms using a

- *Rule-based approach* of manually crafted, i.e., knowledge engineered, rules representing the knowledge of a domain expert, and a
- *Machine learning-based approach* in which advanced mathematical procedures are trained based on a manually labeled dataset.

The article describes the performance of each of the approaches in order to classify the norms of a sub-domain of the German civil law, i.e., tenancy law. Furthermore, an in-depth inspection of the machine learning approach is performed using so-called local-linear model-agnostic explanations (LIME) (Ribeiro et al. 2016a). These explanations allow a detailed analysis of the trained model of a machine learning classifier by approximating the weight input features. Using this insights we cannot only explain the trained behavior or machine learning classifiers, but also use the information to detect overfitting, and to adapt pre-processing to improve the overall text classification. These explanations increase the transparency of the classifier and establish trust with regard to the classification result. In addition, local linear explanations allow us to confirm the hypothesis that the functional type of a semantic norm depends to a large degree on the (modal) verbs and that trained classifiers end up with similar decision structures as knowledge engineered approaches by humans.

The article is structured as follows: Sect. 2 introduces related literature and approaches focusing on the scenario and the technology used to classify legal

norms. Afterwards, in Sect. 3 the legal theory for the classification tasks of legal norms for German statutes is developed and adapted for this empirical task. Section 4 describes the dataset, the rule-based approach, and the usage of supervised machine learning to classify legal norms. A detailed comparison of the accuracy of different classifiers and an extensive discussion of the results are provided in Sect. 4 as well. To explain the results of the machine learning classifier a new technology based on local-linear approximations is used and described in Sect. 5. These explanations can be used to increase the transparency of the trained classifier and increase the interpretability of machine learning for legal text classification.

2 Related work

The computer-assisted analysis of norms in statutory texts with regard to their semantic type and functional role is highly relevant and has attracted researchers ever since. However, hardly any attempt has been made in the German domain, i.e., analyzing German statutes. This is counter-intuitive since the German law documents are well structured. Furthermore, German statutes hardly have any redundancy, at least ideally, and statutes and norms express their meaning in as few words as possible, which makes the semantic density of norms very high. Consequently, their interpretation requires a method and structured schema, which was influenced by logic-based reasoning, which is continuously adapted in legal theory. These factors make them an ideal case for formalization and the analysis with regard to their semantics.

The next two sections describe approaches of classification of legal norms in different jurisdictions. The main differentiation thereby is the technological approach that has been used: Rule-based approaches are described in Sect. 2.1, whereas machine learning-based approaches are described in Sect. 2.2.

2.1 Classifying legal norms with rule-based technologies

Classifying norms within the Dutch legislation has been studied within different scientific projects. In Maat and Winkels (2007) differentiated between different rule types, which they called layers or norms. In their model, they follow the distinction that was already proposed by Hart and Green (2012) and that distinguishes between primary and secondary rules. Primary rules describe the main sources of normative regulation, such as rights and duties. Secondary rules are those norms which regulate the management of rules, such as the applicability and the application scope of norms including transitional provisions.

Based on this model of legal norms, Maat and Winkels (2010) implemented a knowledge engineering, i.e., rule-based, approach by extracting typical text patterns that identify the category of a given norm. The patterns were codified into simple regular expressions and applied to a corpus of 18 Dutch legislative texts (Income Tax Act). In Maat and Winkels (2010, p. 175) also discuss the challenge of finding

the right granularity for classification, and decided to segment a statute on the sentence level. The shortest document consisted only of three sentences, whereas the longest document contained 166 sentences. Finally, they identified and implemented 87 patterns to classify the sentences into 15 different categories. Using this approach they reached an average F_1 score of 0.91. They also analyzed the patterns that they originally determined and how they contributed to the overall result. Their investigation showed that out of the 87 patterns only 44 actually triggered a classification, whereas the remaining ones where not used at all.

A recent approach in determining conflicts in norms was published by Paulo Aires et al. (2017). They used a rule-based approach to distinguish automatically between norm sentences and non-norm sentences. Which is a valid and straightforward differentiation but does not fully reflect the multiple semantic types known in statutory texts. Aires et al. used regular expressions of the form `.+? modal_verb .+` and the modal verbs were manually defined: may, can, must, ought, will, shall. They tested the identification of norms on a dataset consisting of 256 contracts, out of which 92 contracts were manually labeled. The final norm set contained 9864 sentences, i.e., norms. Based on this gold standard their norm identification approach achieved a precision of 0.79 and a recall of 0.98, which leads to the final F_1 score of 0.87.

Wyner et al. used a rule-based approach to extract elements from statutory texts (Wyner and Peters 2010). They selected an excerpt from the US Code of Federal Regulations, US Food and Drug Administration, Department of Health and Human Services.[1] The document contains 1777 tokens, i.e., words, and is relatively small. However, their approach did not focus on the analysis or classification of norms but on their analysis with regard to deontic rules. They performed a very fine granular analysis using the JAPE grammar (Cunningham et al. 2000). Their model covered deontic modals and verbs, agents and themes, and conditional sentences with antecedents and consequences. They reached an average F_1 score of 0.79. However, many of their deontic concepts were extracted without any error ($F_1 = 1$). Wyner et al. extensively discussed potential sources of errors: the main challenge is the syntactic position, and identification, of subject, object, and by-phrases. This becomes even more complex due to the linguistic phenomena of active and passive sentences, which are widely used in German legal documents.

The mentioned approaches focus on the analysis of statutory documents to capture the function and computational semantics of legal norms using rule-based technologies. The next section will follow the same objective but with a different technology, namely using machine learning.

2.2 Classifying legal norms with machine learning-based technologies

Biagioli et al. made an early contribution for the classification of norms in legislative texts using a machine learning approach in 2005 (Biagioli et al. 2005). The

[1] Regulation for blood banks on testing requirements for communicable disease agents in human blood, Title 21 part 610 section 40.

authors differentiate between the "formal partition", i.e., structure, and the semantic units of a regulation. They already identified the potential of semantic annotations as to "make the [information] retrieval easier", which is still a relevant task (Ashley 2017). In order to classify legal norms, the authors distinguish between eleven different semantic types, which are assigned on a paragraph level. For their experiment they used a multi-class support vector machine. The dataset contained 582 manually labeled paragraphs, i.e., sections of a legislative statute. In terms of performance they achieved an average F_1 measure of 0.80, whereas the precision and recall values for different classes where quite diverse, scaling from $F_1 = 0.35$ for the semantic type "permission" to $F_1 = 0.97$ for the semantic type "substitution". Francesconi and Passerini (2007) evaluated the classifiers naive bayes and a support vector machine in determining the same eleven functional types. They showed that the support vector machine performs better in the classification task then naive bayes.

Maat et al. (2010) extended the research done with a knowledge engineering, i.e., rule-based information extraction, by applying a machine learning classifier. This makes it a very interesting and valuable contribution to research, as hardly any attempts exist that make a structured comparison between a knowledge engineered approach and a machine learning based approach. The accuracy rates have again reached the high level of 0.94. They did various different parameter studies and showed that binary term weight, with removal of stop words, and a minimal term frequency of 2 performed best. They did an informal discussion about potential errors and identified the "skewness" as a potential source of errors. This means that classes of norms, that hardly occur throughout legislative texts, tend to be less likely predicted by machine learning classifiers compared to those classes that occur more often.

To the best of our knowledge no attempt has been published on the automated classification of legal norms for the German domain. In addition, only shallow approaches have been made to extend German legal theory, which are not suitable for legal data science, such as automated classification tasks. The next section discusses the conceptual foundation for the automated classification of norms from a theoretical point of view.

3 Semantic types of norms in German statutes

3.1 Legal norms

As law is complex, analytic legal theory aims to disassemble law into its smallest components. Defining those, both legal norms and legal terms are the subject of discussion. As legal terms are often regarded as no more than mere descriptions for legal norms, it is appropriate to concentrate on legal norms as the elementary components of law.

Legal norms can be found in the various sources of law. In a continental legal system as German law, statutory law (German: 'Gesetze') is considered as a main source of law. In our work, we focus on German statutory laws, namely the German tenancy law which is part of the German Civil Code.[2]

3.2 Legal statements

Every statutory text contains legal norms regulating the compliant and non-compliant behavior of those addressed by the rule. There are norms regulating the behavior of respective subjects—'behavior norms'—, as well as norms primarily aiming at regulating decision competencies—'decision norms'. Despite those and other differences, all norms dispose of the following characteristics:

- A universal validity in the sense of a binding behavior requirement or a binding assessment standard its normative character , and
- A claim to be applicable beyond a specific situation, for all situations 'of this kind' within the defined regional and temporal application scope its general character.

Legal rules contained in legal norms have the linguistic form of legal statements. Due to its normative character a 'legal statement' is to be distinguished from a mere 'statement'. A statement usually contains a correlation between an object and a characteristic or a behavior attached to it. This correlation or occurrence is perceived to be of factual or happening nature. As each statement makes the claim that something is or happens in a specific factual way, it is subject to the truth criterion, i.e., it can be attributed the title 'true' or 'false' (Larenz and Canaris 2013, p. 72).

To draw the difference to a 'legal statement', we inspect the third sentence of §535 Abs. 1 BGB: "He [the lessor] must bear all costs to which the leased property is subject".[3] This sentence does not state that landlords (usually) act or are going to act this way. It rather stipulates that all those who are defined as 'landlords' in the meaning of this provision, are obliged to act in a specific way. This is why one cannot ask if this statement is true or false, instead only, if it is a valid component within an effective and consistent legal system.

3.3 Categories

Following a functional classification approach, we have identified nine different categories, i.e., semantic types. Dealing with more categories usually allows for a fine granular distinction, whereas dealing with less categories may present a challenge when it comes to defining the boundaries of functional consistencies.

[2] German Tenancy law is comprised in §§535–597 BGB; to be accessed under https://www.gesetze-im-internet.de/englisch_bgb/index.html.

[3] German: "Der Vermieter hat die auf der Mietsache ruhenden Lasten zu tragen".

Table 1 Semantic types of norms in German civil law statutes

	Semantic type	Description
I	Duty	The primary function of a duty is to stipulate actions, inactions or states
II	Indemnity	The primary function of an indemnity is to clarify that, resp. under which conditions a duty does not exist
III	Permission	The primary function of a permission is to authorize actions, inactions or states
IV	Prohibition	The primary function of a prohibition is to forbid or disallow actions, inactions or states
V	Objection	The primary function of an objection is to define that, resp. under which circumstances an existent claim may not be asserted
VI	Continuation	The primary function of a continuation is to extend or limit the scope of application of a precedent legal statement
VII	Consequence	The primary function of a consequence is to stipulate legal effects, without ordering or allowing character as far as the legal consequence part is concerned
VIII	Definition	The primary function of a definition is to describe and clarify the meaning of a term within the law
IX	Reference	The primary function of a reference is to cite another norm with the aim of total or partial application transfer or non-application

Table 2 Examples of semantic types of norms from the German Civil Code

	Semantic type	Example (German)	Example (English)
I	Duty	Der Mieter ist verpflichtet, dem Vermieter die vereinbarte Miete zu entrichten. (§535 Abs. 1 Sentence 3 BGB)	The lessee is obliged to pay the lessor the agreed rent
II	Indemnity	Veränderungen oder Verschlechterungen der Mietsache, die durch den vertragsgemäen Gebrauch herbeigeführt werden, hat der Mieter nicht zu vertreten. (§538 Abs. 1 Sentence 1 BGB)	The lessee is not responsible for modifications to or deterioration of the leased property brought about by use in conformity with the contract
III	Permission	Die Vertragsparteien können eine andere Anlageform vereinbaren. (§551 Abs. 3 Sentence 2 BGB)	The parties to the contract may agree on another form of investment
IV	Prohibition	Ferner kann der Vermieter sich nicht auf eine Vereinbarung berufen, nach der das Mietverhältnis zum Nachteil des Mieters auflösend bedingt ist. (§572 Abs. 2 Sentence 1 BGB)	In addition, the lessor may not invoke an agreement by which the lease is subject to a condition subsequent to the disadvantage of the lessee
V	Objection	Eine zum Nachteil des Mieters abweichende Vereinbarung ist unwirksam. (§551 Abs. 4 Sentence 1 BGB)	A deviating agreement to the disadvantage of the lessee is ineffective
VI	Continuation	Dies gilt nicht, wenn der Mieter gekündigt hat. (§571 Abs. 1 Sentence 3 BGB)	This does not apply if the lessee has given notice of termination
VII	Consequence	Kennt der Mieter bei Vertragsschluss den Mangel der Mietsache, so stehen ihm die Rechte aus den §§536 und 536a nicht zu. (§536b Abs. 1 Sentence 1 BGB)	If the lessee knows of the defect when entering into the agreement, then he does not have the rights under sections 536 and 536a
VIII	Definition	Ein Mietspiegel ist eine Übersicht über die ortsübliche Vergleichsmiete, soweit die bersicht von der Gemeinde oder von Interessenvertretern der Vermieter und der Mieter gemeinsam erstellt oder anerkannt worden ist. (§558c Abs. 1 Sentence 1 BGB)	A list of representative rents is a table showing the reference rent customary in the locality, if the table has been jointly produced or recognized by the municipality or by representatives of lessors and lessees
IX	Reference	§551 Abs. 3 und 4 gilt entsprechend. (§554a Abs. 2 Sentence 2 BGB)	Section 551 (3) and (4) applies with the necessary modifications

Our aim for legal, linguistic and functional consistency led to the development of the following nine categories (see Table 1): 'Duty', 'Indemnity', 'Permission', 'Prohibition', 'Objection', 'Continuation', 'Consequence', 'Definition' and 'Reference'. Each of those is defined in Table 1 and exemplary sentences are listed in Table 2. From a legal theory perspective, those categories can be divided into complete and incomplete legal statements, which refers to the idea of primary and secondary rules as proposed by Hart and Green (2012). From the nine identified categories only two categories, namely 'definition' and 'reference', can be considered as incomplete legal statements, resp. secondary rules, as they have a supplementary function towards other legal statements.

Duty (I) and Indemnity (II) Deontic logic is an area of logic which deals with normative concepts, systems of norms and normative reasoning. Normative concepts include the concepts of obligation (must, ought to), permission (can, may), prohibition (may not) and related notions. Thus, deontic logic and the theory of normative positions are considerably relevant to legal knowledge representation, and are consequently applied to the analysis of normative systems. According to the 'norm square' by Bentham there are four elementary norm types: 'Duty', 'Prohibition', 'Permission', 'Indemnity', originally named as 'command', 'prohibition', 'non-prohibition', and 'non-command' (Moreso 2014; Hohfeld 1917). A 'Duty' represents an action that must be done; a 'prohibition' an action which must not be done; a 'Permission' indicates an action that can either be performed or not; an 'Indemnity' states that a (required) action does not have to be done. According to this logic, the categories of 'Duty' and 'Indemnity' on the one hand and 'Permission' and 'Prohibition' on the other hand are considered to be in contradictory relationship. Consequently , the categories of 'Duty' and 'Prohibition' could be considered parts of the overall category 'Obligation', and the categories of 'Permission' and 'Indemnity' part of the overall category 'rights'. In this work we have, however, considered the categories separately. From a linguistic perspective, there is a closer relationship be tween the categories of permission and prohibition on the one hand and duty and indemnity on the other hand, whereby each latter category is the negative variation of each former category.
The category 'Duty' containing a statutory obligation to act in a specific manner is usually characterized by terms as 'hat zu tragen' (shall carry), 'soll tragen' (shall carry) and the equivalent 'ist verpflichtet zu tragen' (is obliged to carry). The words 'sollen' (shall) and 'verpflichtet sein' (is obliged to) are perceived as the key words of the normative language. The 'Indemnity' category, linguistically characterized by enriching the 'Duty' category by adding negative elements such as 'nicht/kein' (not/none) can also emerge as a negation of the first one.
Permission (III) and Prohibition (IV) As mentioned before, the categories of permission and prohibition are in a contradictory relationship, where something that is prohibited cannot be allowed and something what is allowed cannot be prohibited. Whereas a permission represents a statutory granted entitlement or legitimation, a prohibition as an order to abstain from action is regarded as the equivalent of the negation of a per-mission. Linguistically, a permission can be identified by

using expressions such as ‘kann/darf/ist berechtigt’, which are combined with the negative terms of ‘nicht/kein’ in case of the ‘Prohibition’ category.

Objection (V) An ‘Objection’ is a legal statement expressing a means of defense in material law against the assertion of a claim (German: ‘Einwendung’). In contrast to a ‘Consequence’, an ‘Objection’ aims for setting a specific, not general, stipulation into effect, namely that a claim may not be asserted. It is characterized by terms as ‘ist unwirksam’ (is ineffective) or ‘ist unzulässig’ (is inadmissible).

Continuation (VI) This category is characterized by two regulatory techniques that the German legislator uses regularly: On the one hand, in those cases in which the legal consequence in a preceding and subsequent legal statement would be the same, it is not repeated, but only referred to in the latter one. This kind of technique, where the legal consequences of preceding statements are fully or partially transferred into subsequent legal statements, is characterized by formulations like ‘the same applies’ in the subsequent legal statements.

On the other hand, as legal statements are often formulated quite broadly, they need to be restricted by subsequent negative legal statements. Those restrictive legal statements are characterized by containing negative validity orders, using formulations like ‘not applicable’ with respect to a preceding positive legal statement. The reason behind these legal constructions is firstly, that the inclusion of all restrictive elements into a positive legal statement would lead to clumsy and incomprehensible sentences. Secondly, the German Civil Code is dominated by the tendency to work according to a ‘rule’/‘exception’ pattern mechanism and thereby regulating implicitly the question of burden of proof.

Consequence (VII) When defining the category of ‘Consequence’ it is useful to contrast it from a duty. A ‘Duty’, on the one hand, as an order to do something addresses specific addressees and requires a certain behavior. A ‘Consequence’, on the other hand, does not necessarily make a behavior of someone to the object of its regulation; instead it concentrates on setting a legal stipulation into effect. A ‘Duty’ aims for obedience, a ‘Consequence’ for validity. Whereas the direct effect of a ‘Duty’ happens at the factual level, the direct effect of a ‘Consequence’ takes place at a normative level. The category of ‘Consequence’ is characterized by use of finite verbs in the third person and the simultaneous absence of any verb forms characteristic for any of the above mentioned categories. Example: §542 Abs. 2 BGB: ‘A rental agreement that has been received for a certain period ends (= finite verb in 3. person) with the expiration of this period (= absence of any modal (and infinite) verb)’.

Definition (VIII) Explanatory legal statements as secondary rules are either defining a term used in other legal statements or specifying the content of a general term with regard to different situations. Whereas defining legal statements usually relates to elements of the offense, e.g. BGB §90: ‘Sache’, §932: ‘guter Glaube’, §276: ‘Fahrlässigkeit’, specifying legal statements relates to the legal consequence part, e.g. §§249 pp. BGB. The ‘Definition’ category comprises the defining legal statements which are characterized by formulations as ‘liegt vor, wenn’ (is given if) or ‘im Sinne des Gesetzes’ (as defined by law). As German law uses legal fictions where it equally could have used the instrument of definitions, legal fictions are also considered as elements of this category. A legal fiction presents

an intended equalization of evidently unequal situations, respectively terms. The fundamental difference to a subsumption error is the fact of awareness of the factual inequality. One could argue that statutory fictions intend to transfer a rule contained in one legal statement to another one, and therefore have the role of hidden references. However, as the prime aim of fictions is to order that certain factual or legal facts be regarded as factually given and cannot be rebutted in legal proceedings as they constitute by definition a deviation from the factual situation, they bear a strong functional proximity to 'Definitions'. Linguistically they are characterized by terms as 'gilt als' (be classified as, to be treated as).
Reference (IX) A 'Reference' as a legal statement cites another norm, resp. part of it, with the aim of direct or analogous application of this norm. The 'Reference' can thereby refer to the legal consequence part of a norm only, or to the norm as a whole. It is characterized by terms as 'ist anzuwenden' ('applies') in connection with concrete norm articles.

Based on these considerations of a functional classification of legal norms Table 2 provides concrete examples for each category. The next section is dedicated to outline the difference between the function of a norm and its content and discusses the classification patterns for the functional classification.

3.4 Function and content of legal norms

One can look at a legal statement from a content-based or a functional perspective. §535 Abs. 1 Sentence 2 BGB: The landlord must surrender the rental object to the tenant in a condition suitable for use in conformity with the contract and maintain it in this condition for the rental period.

From a functional perspective, this legal statement contains one duty—the 'must' statement—, from a content-based perspective, two duties—the duty 'to surrender' and another one 'to maintain'. The content-based approach can be seen as specifying the functional findings, respectively the functional part can be seen as having a general function with the content-based part having a completing or specifying function. Both categories follow different linguistic patterns, as described in the following.

3.4.1 The role of modal verbs in defining the semantic type norms

Law has factual, normative and linguistic dimensions. Indeed, our understanding of law is more a matter of the interpretation of literary texts than of the scientific description of physical objects. One of the common interpretation methods in law is the 'grammatical interpretation', stressing the word sense of the legal statement. Legal statements are formed by sentences, respectively their syntactical elements: nouns, adjectives and verbs. Neither nouns nor adjectives have been identified as being dominant in defining functional relations; in contrast to those, verbs are intended to express an activity, an event or a state and therefore capable of capturing the interactions between different sentence elements.

Considering functional analysis, 'modal verbs', as a labeling for verbs which are expressing a modality, play a central role in German language. In German, the verbs 'dürfen'/ 'mögen' (may), 'können' (can), 'müssen' (must), 'sollen' (shall) and 'wollen' (want) are considered part of this category. As they often have an auxiliary function towards other verbs, they are often labeled 'modal auxiliaries'. The specific attribute of modal verbs is their capability to characterize statements, resp. other verbs. Thus, they are able to add a new functional dimension to the sentence, which goes beyond that one contributed by non-modal verbs. From a deontic perspective, we can identify deontic necessities as 'The tenant must pay the rental fee', and deontic possibilities as 'The tenant may use the flat'.

Legal statements are usually composed of a 'factual' offence part and a 'legal' consequence part. A legal statement is an expression of the link between the abstractly described facts of the case—the 'offence'—and the equally abstractly described 'legal consequence'. In every situation where the facts laid down in the legal statement are fulfilled, the legal consequence is set into effect. It is the legal consequence part which makes the necessary normative link and creates an 'order of validity'. As modal verbs are usually used to express the legal consequence part, they naturally dispose of a strong connection to the normative value contained therein, and a significant value for any normative functional classification. Based on these considerations, we have identified the following modal verbs as being central for the classification tasks as described within this article (see Table 2):

Duty modal verbs 'must', 'have to', 'shall' have a high characterizing value;
Permission modal verbs 'can', and 'may' have a high characterizing value;
Prohibition or Indemnity the negated form of the verbs above, e.g. 'shall and shall not', 'must and must not' have been identified

3.4.2 The general role of verbs in classification tasks

The dominant function of modal verbs as finite verbs in functional classification reflects the important role of verbs in classifications in general. Verbs can be generally classified into the categories of finite, e.g. '(er) überträgt' (he transfers), and infinite verbs, e.g. 'übertragen' [(to) transfer]. Finite verbs are verb forms which are subject to inflection; this is why they are capable of expressing certain grammatical attributes (e.g. person, numerus, tempus, modus) and therefore relate to other elements of the sentence, as e.g. the (nominative) subject. It is their inflective character which allows them to transmit functional value with respect to legal statements. Finite verbs can occur in the form of main verbs which do not require any additional verbs besides, or auxiliary verbs, e.g. modal verbs.

In contrast, infinite verbs are not subject to inflection, so that their proximity to other parts of the sentence is limited. Instead, they may be found in the role of defining, respectively specifying the content-dimension of a functionally pre-defined sentence (see Sect. 3.4.1). As the 'verb besides the (finite) verb', their main point of reference is the finite verb; thus, their primary role might be seen in a content-related function with respect to this verb (and the functions thereof).

With regard to a rule-based approach, not only the role of verbs, but also their location is crucial. Considering legal language, sentences are often composed of main and subordinate sentences. This method is used to structure different parts of legal statements and thus reduce their complexity. We have seen above, that modal verbs as finite verbs usually occur in the legal consequence part of a sentence. We have also observed that the legal consequence part of a legal statement is usually located in the 'main' part of the sentence, whereas the factual offence part is to be found in the subordinate sentence part. As the legal consequence is the part which makes the necessary link for the transformation from a statement to a legal statement, it is the main sentence part which is most relevant for a normative functional classification. With legal theoretical and linguistic considerations, the classification tasks were performed using two different and complementary technologies, namely rule-based classification and machine-learning based classifications, which are described in the next section.

4 Classification of legal norms with natural language processing

4.1 Dataset

The used dataset comprises 601 sentences which constitute the tenancy law of the German Civil Code (§535–§597) in its consolidated version effective from 21st February 2017.

In a first preprocessing step, the raw text of these articles was segmented into sentences. As sentence boundaries sentence ending periods were chosen. While sentence segmentation—especially in the legal domain (Šavelka and Ashley 2016)—can be a challenging task, a straight-forward approach was chosen. Remaining punctuation marks, i.e., semicolons, colons, comma, and periods, following abbreviations or in other, non sentence-ending position were ignored. In the case of enumerations, the same rules are applied. This implies that all items in an enumeration are considered as a single sentence unless one of them ends with a period. In this case, the next sentence starts with the following enumeration item. This approach works well within the chosen domain, as the appearing enumerations almost all (except one) embody a single sentence. This method is not suitable for general sentence segmentation tasks, where enumerations can contain multiple nested sentences. The item labels (e.g. "a." or "1") were not stripped from the text. The subject of classification, i.e., sentences, were chosen as they allow for a consistent segmentation and ideally encode exactly one legal statement (Bundesministerium der Justiz 2008) (see also Sect. 3).

In the next step, the 601 sentences were manually classified by a single domain expert, according to the taxonomy described in Sect. 3. Table 3 shows the distribution of the different semantic types. Some types occur regularly, e.g. III 'Permission', VII 'Consequence', and I 'Duty', whereas some have very low support, e.g. II 'Indemnity', VIII 'Definition', IV 'Prohibition', and VI 'Continuation'.

Table 3 Manually labeled dataset consisting of sentences extracted from the German tenancy law

Semantic type		Occurrences	Rel occurr. (%)
I	Duty	117	19
II	Indemnity	8	1
III	Permission	148	25
IV	Prohibition	18	3
V	Objection	98	16
VI	Continuation	21	3
VII	Consequence	117	19
VIII	Definition	18	3
IX	Reference	56	9
		Σ 601	100

4.2 Rule-based classification of legal norms

For the classification of the sentences in the dataset using rule-based classification, two main approaches can be used: standard regular expressions for each category, or more sophisticated rule languages, supporting a complex type system, such as UIMA Ruta (Klügl 2014) or JAPE (Cunningham et al. 2000). While regular expressions are a powerful tool in information extraction despite their simplicity, dedicated rule languages have major advantages regarding this classification task, such as the ability to work with type systems that define higher-level entities such as verbs or nouns. For example, it is not viable to write a regular expression that considers the information whether a sentence contains a verb, as this information is simply not available in the raw text.

As the workbench used to perform the preprocessing steps is built on the UIMA framework (Waltl et al. 2016), the rules for classification are implemented in the UIMA Ruta language, using a single script for each category. The workbench first creates all necessary text information such as Part-of-Speech tagging, which is then consumed by the script performing the actual classification.

By consuming the information provided by the Part-of-Speech component (lines 2–5), phrases indicating a duty are extracted (lines 8–13, e.g. “ist zu entrichten”, engl. “is to be paid”). Lines 1 and 7 are declarations of annotation types later used for creating the final, classifying rule (line 14). In general, the UIMA Ruta scripts used for this classification comprise between 20 and 80 lines of code, with between 4 and 15 final rules for each category.

Listing 1 A snippet of a script, extracting infinitive phrases in sentences.

```
DECLARE INFINITIV;
V.PosValue == "VVIZU" {-> INFINITIV};
V.PosValue == "VAINF" {-> INFINITIV};
V.PosValue == "VNINF" {-> INFINITIV};
V.PosValue == "VVINF" {-> INFINITIV};

DECLARE ISTINFINITV;
(W{REGEXP("ist|sind")} # W{REGEXP("zu.*")} INFINITIV) {
-> ISTINFINITV
};
(W{REGEXP("hat|haben")} # W{REGEXP("zu.*")} INFINITIV) {
-> ISTINFINITV
};
Sentence{CONTAINS(ISTINFINITV) -> Duty};
```

Using this setup, four iterations were performed in which the rules were adjusted based on the results and evaluation of the previous iteration. An overview of all iterations can be seen in Table 4. Thereby, the different semantic types are differentiated and the precision and recall is determined for every type individually. The table shows, that the highest F_1 score was achieved in the last iteration, with the exception of category IX 'Reference'. Looking at the trends of precision and recall for each category, the advantages of rules to independent adjustment and improvement are directly indicated by each metric. Type III 'Permission' serves as an example: while the precision of this class was reasonable satisfying in the first iteration, recall was rather low. In the subsequent iterations, the recall was increased to an acceptable level, while precision was almost stable through all iterations. Both measures can obviously be increased simultaneously, as seen at type VII 'Consequence'.

When integrating the results of the evaluation into the next generation of the scripts, one runs the risk of integrating edge cases and single occurrences of patterns that have been unveiled by the evaluation. To mitigate this kind of ad-hoc fitting and metric boosting, the adjustment was restricted on the rule level for example by improving the range of repetition quantifiers (e.g. star repetition to upper bounded repetition count), allowing arbitrary amounts of token within parts of the rules, or restricting the types of tokens (e.g. instead of matching any tokens, only considered verbs in infinitive). In a few cases, new rules were added that capture a whole new group of sentences, not individual ones. Another measure taken in each iteration was the elimination of implementation bugs within the rules.

While the overall quality of the rules is acceptable, category IV 'Prohibition' and VIII 'Definition' perform unsatisfactorily. For these categories and in general, an overall better scoring of the rules could be performed if the sentence segmenting module would not only create whole sentences but also auxiliary sentences. As German legal texts make heavy use of long, nested sentences this information would be beneficial for the implementation of the rules. Many cases were observed in

Table 4 Four iterations of rule-based norm classification in German tenancy law

Semantic Type			Iterations			
			I	II	III	IV
I	Duty	Precision	**0.673**	0.658	0.630	0.634
		Recall	0.497	0.626	0.839	**0.839**
		F1	0.571	0.642	0.720	**0.722**
II	Indemnity	Precision	0.194	0.194	**0.715**	0.714
		Recall	0.375	0.375	0.385	**0.385**
		F1	0.255	0.255	0.500	**0.500**
III	Permission	Precision	**0.886**	0.854	0.822	0.822
		Recall	0.531	0.530	0.831	**0.831**
		F1	0.664	0.654	0.827	**0.827**
IV	Prohibition	Precision	0.327	0.286	0.857	**0.857**
		Recall	**0.500**	0.100	0.316	0.316
		F1	0.395	0.148	0.462	**0.462**
V	Objection	Precision	0.895	**1.000**	0.990	0.983
		Recall	0.298	0.048	0.893	**0.922**
		F1	0.447	0.091	0.939	**0.951**
VI	Continuation	Precision	0.947	0.947	0.947	**0.950**
		Recall	0.514	0.545	0.600	**0.633**
		F1	0.667	0.692	0.735	**0.760**
VII	Consequence	Precision	0.406	0.242	0.824	**0.832**
		Recall	0.211	0.238	0.748	**0.748**
		F1	0.278	0.240	0.784	**0.788**
VIII	Definition	Precision	0.146	0.127	0.157	**0.295**
		Recall	0.250	0.400	0.381	**0.520**
		F1	0.185	0.193	0.222	**0.377**
IX	Reference	Precision	0.783	0.833	0.833	**0.833**
		Recall	0.771	**0.873**	0.696	0.696
		F1	0.777	**0.853**	0.759	0.759
Arithmetic mean (weighted)		Precision	0.697	0.674	0.798	**0.803**
		Recall	0.435	0.427	0.771	**0.781**
		F1	0.518	0.465	0.773	**0.782**

Best value for each semantic type highlighted

which the existence of auxiliary sentences caused severe problems for a rule-based approach to identify the right pattern.

An example can be seen in §556 Abs. 3 Satz 3 BGB: 'After expiry of this period, the assertion of a claim by the lessor is excluded, unless the landlord is not to be held responsible for the late assertion'.[4] Whereas in the main sentence the words

[4] German: 'Nach Ablauf dieser Frist ist die Geltendmachung einer Nachforderung durch den Vermieter ausgeschlossen, es sei denn, der Vermieter hat die versptete Geltendmachung nicht zu vertreten'.

'is excluded' indicate the category 'objection', the formulation 'is not to be held responsible' contained in the negative if-sentence as auxiliary sentence beginning with 'es sei denn', is characteristic for the category 'indemnity'. A separation of main and auxiliary sentence would exclude those multi-functionalities and thus most probably lead to clearer classification.

However, the detection of auxiliary sentences in the German language—especially in environments with high sentence complexity—is an error prone process. Thus, more complex rules using finer-grained linguistic information are subject to future work. Another source of error for all types of scripts was the excessive use of unbounded wildcard elements ('*' in regular expressions) and sometimes the explicit ordering of rule elements. For example, if a pattern was specified as 'T followed by V' because all of the samples this pattern was based on, contained 'T' and 'V' in that order, but in other samples, a different sequence occurs, the rule would not trigger on them. Other linguistic subtleties cause to fail rules as well, which leads to many "near misses".

Interestingly, the types with the lowest frequency in the dataset, i.e. II 'Indemnity', IV 'Prohibition', VIII 'Definition', and VII 'Continuation', also have the lowest F_1 score using these rule based approaches. Continuation however performs much better than any of these categories. We assume that a possible reason might be seen in the differences concerning the average length and complexity of the sentences between the categories, with the category of 'continuation' containing on average shorter and simpler sentence structures. Also, the categories of 'indemnity' and 'prohibition' are characterized by negative statements which might add up to complexity.

The comparably low scoring for the category of 'definition' might be due to the fact that definitions contained in German civil law tend to be embedded in more complex sentence structures and thus occur at the intersection of at least two, if not three different categories. As our classification system assigns one sentence with one category only, according to the main focus of the sentence, this can lead to a neglect of subordinate functions of the sentence which results in low scoring result.

Another anomaly is the collapse of recall of categories IV 'Prohibition' and V 'Objection' in iteration two. These outliers stem from overly aggressive adjusting of rules. This illustrates how careful revision based on testing results can result in an increased quality of the rules, but also that these changes have to be selected carefully.

4.3 Supervised machine learning to classify legal norms

4.3.1 The KDD process

The implementation of the classification of legal norms using supervised machine learning followed a basic workflow consisting of the following steps (illustrated in Fig. 1):

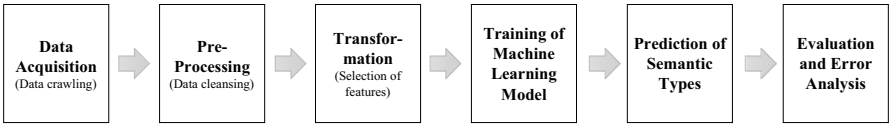

Fig. 1 The process steps for the classification of legal norms

Data acquisition As train and test dataset we used the same 601 labeled sentences which is described in Sect. 4.
Pre-processing Where indicated stopwords, according to the NLTK German stopword-list,[5] have been removed from the sentences. Apart from this, no further pre-processing has been applied.
Transformation (incl. feature selection) As features, bag-of-words, word count vectors of the sentences have been used. Where indicated an additional tf-idf transformer has been applied on these vectors.
Training of machine learning model 5 different classifiers were applied on the task of predicting the semantic types of German legal norms. The models were trained with 80% of the dataset (480 samples) using a tenfold cross-validation.
Evaluation and error analysis To evaluate the performance of the trained models, weighted variants of precision, recall and F_1 were used. In addition to such standard evaluation metrics, the local linear explanations for the models were extracted using the LIME library (Ribeiro et al. 2016b). Based on these results, an attempt has been made to reconstruct the decisions and errors made in the classification process.

4.3.2 Prediction of semantic types

As classifiers we used multinomial Naive Bayes (MNB), logistic regression (LR), support vector machines (SVM) with linear kernel, random forests (RF) and multilayer perceptrons (P), each in combination with the transformations described in Sect. 4.3.1. This results in 15 different configurations (see Table 5). The training, application and evaluation of the models were implemented in Python using the scikit-learn library (Pedregosa et al. 2011). In addition to the standard evaluation, the local linear explanations were extracted for the SVM classifier using the LIME library (Ribeiro et al. 2016b).

4.3.3 Comparison of classifier performance

The results of the trained classifiers vary heavily between classifiers and classes. The overall (weighted) average evaluation of the classifiers' performance in Table 5 shows a minimum F_1 score of 0.57 (MNB + SW + TF-IDF) and a maximum score of 0.83 (SVC).

[5] http://www.nltk.org/book/ch02.html, accessed on 03/19/2018.

Table 5 Evaluation of five common classifiers (SVC: Support Vector Classifier (linear kernel); RF: Random Forest; P: Multilayer Perceptron; MNB: Multinomial Naive Bayes; and LR: Logistic Regression) using bag-of-words and three different input feature pipelines: stopword removal and tf-idf calculation for tokens

ML classifier (SW: stopword removal) (TF-IDF: weighting of tokens)		Precision (avg; std)	Recall (avg; std)	F1(avg; std)
SVC		**0.85** (± 0.09)	**0.84** (± 0.07)	**0.83** (± 0.08)
	+ SW	0.76 (± 0.14)	0.76 (± 0.11)	0.75 (± 0.12)
	+ SW + TF-IDF	0.74 (± 0.12)	0.74 (± 0.12)	0.72 (± 0.11)
RF		0.71 (± 0.16)	0.73 (± 0.17)	0.72 (± 0.15)
	+ SW	0.69 (± 0.11)	0.67 (± 0.14)	0.67 (± 0.11)
	+ SW + TF-IDF	0.68 (± 0.12)	0.66 (± 0.12)	0.66 (± 0.16)
P		0.80 (± 0.09)	0.77 (± 0.10)	0.76 (± 0.09)
	+ SW	0.70 (± 0.11)	0.68 (± 0.10)	0.67 (± 0.11)
	+ SW + TF-IDF	0.73 (± 0.12)	0.70 (± 0.12)	0.70 (± 0.12)
MNB		0.68 (± 0.13)	0.70 (± 0.12)	0.66 (± 0.12)
	+ SW	0.63 (± 0.14)	0.65 (± 0.11)	0.62 (± 0.11)
	+ SW + TF-IDF	0.64 (± 0.08)	0.60 (± 0.11)	0.57 (± 0.10)
LR		0.82 (± 0.10)	0.82 (± 0.09)	0.81 (± 0.09)
	+ SW	0.75 (± 0.12)	0.75 (± 0.09)	0.73 (± 0.09)
	+ SW + TF-IDF	0.66 (± 0.12)	0.65 (± 0.11)	0.63 (± 0.10)

Weighted average and standard deviation for each classifier. Best performances bold face

Other configurations vary between around 0.6 and 0.7. Compared to the rule-based classification, most of the classifiers (with the exception of SVC and LR) perform worse in terms of F_1 score. Also remarkable is that all classifiers perform best when no transformation of the input data is applied, i.e., no stopword removal or tf-idf calculation. Even if sometimes only by a small margin, rules perform better than most of the tested classifiers in different feature pipelines. In an attempt to explain this difference, a possible source for the better performance of rules is the size and the homogeneity of the dataset. While the absence of a satisfiable amount of samples for some classes (e.g. Indemnity, Prohibition, Definition) is not necessarily better mitigated with rules in our case (see Table 4 for performance of rules), the sizes of samples for a specific feature phenomenon might be. With rules and the help of a domain expert, we can quickly adapt to corner cases or outliers in a specific class within the dataset. Classifiers however, might not be sufficiently trained with only a few samples of such a phenomenon. With a more extensive dataset, we expect other classifiers than the SVC to also surpass the performance of our rigid rules. A more in-depth attempt to draw a connection between a specific classifier's predictions and handcrafted rules is given in Sect. 5.

Table 6 Inspection of the performance differentiated by semantic types of the Support Vector Classifier using a linear kernel

Semantic Types		Precision	Recall	F1	Support
I	Duty	0.92	0.96	0.94	24
II	Indemnity	0.50	0.50	0.50	2
III	Permission	0.94	1.00	0.97	31
IV	Prohibition	0.75	0.75	0.75	4
V	Objection	0.94	0.84	0.89	19
VI	Continuation	1.00	1.00	1.00	3
VII	Consequence	1.00	0.84	0.91	25
VIII	Definition	0.33	1.00	0.50	1
IX	Reference	0.92	1.00	0.96	12
Arithmetic mean (weighted)		0.93	0.92	0.92	121

4.3.4 Error analysis

To be able to better understand the classification process of the selected models, the best configuration (SVC) is examined in greater detail (Table 6). Similar to the rule-based classification, precision, recall and F_1 vary heavily in and between the classes, especially in those with very low frequency (e.g. VIII "Definition", II "Indemnity").

For the classes with higher support (e.g. Permission, Consequence) the performance is much higher compared to the rule-based classification. The mean performance of the SVC classifier shown in Table 6 is different from the performance of the same classifier in Table 5. This is because of different train test splits applied. While Table 5 is generated from a tenfold cross-validation method, Table 6 is based on a static test split. In both cases 80% of the dataset was used as training data. The weighted mean mitigates the positive impact of small classes such as Definition, where only one instance is present in the test set.

In the end, the overall mean performance is however better than the results of the rule-based classification. As the same base dataset was used to train and evaluate the two approaches, the reason for this is not based on the quality of the data, but only the classification type. An attempt to explain the difference between the two approaches and their reasoning behind the classification is given in the next chapters.

5 Explaining the classification of supervised machine learning

Explaining the behavior of a rule-based information extraction component is well understood and can be done with reasonable effort. Tracing the rules and determining whether they apply or not with regard to a certain pattern is possible. Also the determination of so-called 'near misses' can be done. Near misses describe textual patterns that are very close to the formal description of a rule but do not match because of a small deviation. However, in contrary to heuristic machine learning based approaches rules either match or do not match, there is no confidence other than 0 or 1. Consequently, rule-based approaches are explainable and their results can be reproduced and understood by humans (Chiticariu et al. 2013).

This does in general not apply for machine learning based approaches and trained linguistic models (Waltl and Vogl 2018). A predominant and commonly accepted understanding is that the behavior of machine learning classifiers can only be observed like a black-box, which produces deterministic predictions for a given input. However, recent advances in machine learning allow to inspect the reasons for a prediction using model-agnostic explanation technology based on local linear approximations (Ribeiro et al. 2016a).

5.1 Local linear approximations and their role for machine learning

In order to explain the characteristics of a machine learning classifier so-called explanation methods have been created. These methods are attempts of providing interpretable explanations, which is defined by Ribeiro et al. (2016a) as the provision of a qualitative understanding between the input variables and the response. As there are multiple approaches that contribute to this, recent research focused on local fidelity, which concentrates on examining how a trained machine learning model behaves in the vicinity of the instance being predicted. This means that is not representative for the complete prediction model but for single instances.

Ribeiro et al. (2016a) published an approach, called LIME (local interpretable model-agnostic explanations), which provides a model-agnostic methodology to assess any machine learning classifier on an instances level to provide a human interpretable representation of an explanation for the behavior of a classifier. We adapted their approach for the classification tasks for legal norms to inspect the decisions made by the trained models from our dataset and received remarkable insights, which could be used to further improve the classifier, to validate the reliability of the classifier and to increase the trust in machine learning approaches for text classification in general.

5.2 Analysis of the trained classifier for legal norm classification

Based on the best classifier for the legal norm classification, namely the Support Vector Classifier with a linear kernel (see Sect. 4.3.3), we used LIME to create explanations for the classification task on the instance, i.e., sentence level. The results for the classification are illustrated in two tables, namely Tables 7 and 8.

Table 7 shows the classified sentence (first row), which is a sentence from the German tenancy law §588 German Civil Code.[6] The official translation of the norm is as follows: 'The usufructuary lessor must compensate the lessee for expenses incurred and earnings lost as a result of the measure to an extent appropriate to the circumstances'. The classifier predicts that the norm is a *Duty* with a comparably high confidence of 0.80. This is also the true class with regard to the manually classified dataset. The table shows the five tokens, i.e. words, that positively contributed to the *Duty* classification. In addition, also the tokens with a negative weight are

[6] https://www.gesetze-im-internet.de/englisch_bgb/englisch_bgb.html#p2426.

Table 7 Explanation of the classification of a duty norm from German tenancy law

Der Verpächter hat die dem Pächter durch die Maßnahme entstandenen Aufwendungen und entgangenen Erträge in einem den Umständen nach angemessenen Umfang zu ersetzen			
Predicted class		Duty (Confidence: 0.80)	
True class		Duty	
Five tokens (pos. weight)		Five tokens (neg. weight)	
hat	0.23	Der	− 0.03
zu	0.17	angemessenen	− 0.03
Pächter	0.09	Aufwendungen	− 0.03
dem	0.09	durch	− 0.07
ersetzen	0.08	den	− 0.09

Table 8 Explanation of the classification of a reference norm from German tenancy law

(4) Auf das dem Mieter nach Absatz 2 Nr. 1 zustehende Kündigungsrecht sind die §§536b und 536d entsprechend anzuwenden			
Predicted class		Reference (Confidence: 0.94)	
True class		Reference	
Five tokens (pos. weight)		Five tokens (neg. weight)	
entsprechend	0.51	die	− 0.01
anzuwenden	0.24	das	− 0.01
sind	0.08	Mieter	− 0.02
und	0.06	dem	− 0.05
Absatz	0.05	nach	− 0.06

displayed. The weight of the tokens are estimators for the contribution the overall confidence of a classifier with regard to a classification, e.g. *Duty*. Interestingly, the tokens 'hat' (engl. 'has'), 'zu' (engl. 'to'), and 'ersetzen' (engl. 'compensate') are identified by the machine learning classifier to be most valuable during the classification task. This is a remarkable result as "hat zu ersetzen" forms a valid verb phrase, which is identified by the classifier. Based on the input that we received from our legal expert this corresponds to the classification as it would be performed by humans and it is in line with theory that the (modal) verbs significantly determine the functional semantic type of a norm. The tokens that are identified to have a negative impact consist of three articles or prepositions ('Der', 'durch', and 'den'). However, it can be seen that the weight of the tokens are small compared to the weight of the tokens that positively contribute to the classification.

Another example for the detailed inspection of a legal norm of type *Reference* is shown in Table 8. The table follows the same structure as Table 7 but the classified norm is from §543 German Civil Code:[7] 'Sections 536b and 536d are to be applied with the necessary modifications to the right to notice of termination to which the lessee is entitled under subsection (2) no. 1'. The classifier's confidence is 0.94, which is a remarkably high value. The five tokens, which contributed most to this classification are the tokens 'entsprechend' (engl. 'with the necessary modifications'), 'anzuwenden' (engl. 'applied'), and 'sind' (engl. 'are to be'). Again, the legal expert confirmed that the most significant tokens reflect the classification as it would be performed by a human. The tokens with a negative weight mainly consist of stop-words and have only a minor impact on the classifiers confidence as its confidence is high.

Based on this example the behaviour of the trained machine learning model with regard to the classification on an instance level can be validated. This allows an inspection of how the classifier draws his conclusions and in which cases one could (not) trust the classifier. Ultimately, this increases the trust of machine learning and demystifies the functioning of machine learning in terms of classification and prediction tasks.

The next section compares the internal decision structure of the trained machine learning classifier with the handcrafted rules from Sect. 4.2 and illustrates similarities and deviations between the machine learning and the knowledge engineering approach.

5.3 Comparison of knowledge engineering and machine learning model

In order to allow for a comprehensive comparison between handcrafted rules, i.e. knowledge engineering approach, and the trained machine learning model the weight of the individual tokens was assessed. During the structured analysis in which every instance, i.e., sentence, was inspected the weight of each token for each

[7] https://www.gesetze-im-internet.de/englisch_bgb/englisch_bgb.html#p1978.

Table 9 Top and bottom tokens for each semantic type determined and based on the aggregation of weights for each instance (note that the weights can also be negative)

Token weights	Semantic types								
	I	II	III	IV	V	VI	VII	VIII	IX
Top 5 token	muss	nicht	kann	kann	ist	gilt	berücksichtigt	von	entspre
	sind	zu	können	berechtigt	ausgeschlossen	Dies	nicht	der	anzuwenden
	hat	oder	darf	der	unwirksam	Dasselbe	erste	oder	gelten
	zu	herbeigeführt	berechtigt	in	zulässig	fuer	faellig	Gemeinde	Satz
	ist	Verschlechterung	vereinbaren	Mieter	Vereinbarung	wenn	ihm	und	gilt
Bottom 5 token	soweit	Mietsache	Während	Vereinbarung	mit	ueber	sind	werden	des
	entspre	hat	an	sich	Satz	der	Kündigungsschr	einen	dem
	nicht	den	Sind	soll	Absatz	die	eine	auf	werden
	den	durch	Satz	zum	Pacht	ist	gilt	fuer	der
	das	die	nicht	nicht	der	auf	ist	die	im

classification was determined. Again the machine learning classifier with the highest overall accuracy was used, i.e. support vector machine with a linear kernel.

The result is a table in which each record represents a token in a norm and its contribution to the confidence of the classification. To avoid biases, only correct classifications were used. The table can be aggregated according to the semantic types and the tokens are ordered with regard to their accumulated weight, i.e., sum. The five tokens with the highest and the lowest aggregated contribution to the classifier's confidence are shown in Table 9.

Table 9 shows the nine semantic types (columns) that are differentiated by the classifier and the corresponding tokens that are decisive for each of the individual classes. We compared these automatically created lists with the results from the rule-based approach as described in Sect. 4.2 and found significant overlaps in the tokens used for the classification using machine learning, i.e. Support Vector Classifier, and the rule-based information extraction.

It is remarkable, that the most decisive tokens for the classification of a Duty ($F_1 = 0.94$) are verb phrases consisting of 'muss', 'hat', 'sind' in combination with 'zu' (correspondence in English: has to). This reflects to a large degree the knowledge as explified in rules by the domain expert. Listing 1 The decisive tokens for the class Permission ($F_1 = 0.97$) follow the same principle: 'kann', 'können', 'darf', 'berechtigt'. These correspond to the English verbs such as 'can', 'allowed', 'permitted'. The tokens with a low or negative weight consist to a large extent of stopwords, mainly pronouns and determiners. Interestingly, the term 'nicht' (engl. not) occurs in the negative weight column for a Prohibition (IV). This stands in contradiction to the rule-based approach where 'nicht' is considered a positive weight factor. The Tables 7, 8, and 9 allows also an indication why the ML setups, in which stop-words have not been removed, perform better. It seems as if the negative weight of stop-words helps the classifiers to differentiate between classes. However, this hypothesis is hard to verify solely on the data at hand as it would require an explicit in-depth experiment.

A structured experiment to compare the results from the local-linear explanations is missing. However, the results at hand look promising to understand the trained classifier and to assume that it—at least partially—reflects the decision structure of a domain expert.

5.4 Future research and potentials of local-linear explanations

The usage of local-linear explanations to assess the function and malfunction of a machine learning classifier for text classification offers a lot of potential and opportunities. Since the analysis of trained classifiers has been a challenge the LIME technology can help to interpret, to improve, and finally to understand a trained classifier. Using local-linear approximations the internal structure is—at least partially—unveiled. In addition, this contributes to a demystification of the trained classifier and machine learning in general.

Based on our insights we see at least five research directions that could benefit from this method in the field of legal text classification:

1. *Pre-processing* Using detailed information on the classification for a single document or sentences pre-processing steps, such as stopword removal, could be adapted to create tailored stopword lists that remove words not necessary for the classification task.
2. *Feature selection and weighting* The analysis of the contribution of particular features with regard to the classification could be used to adapt the feature selection, e.g. bag-of-words, bag-of-verbs, n-grams, etc., and analyze the performance on the instance level.
3. *Parameter optimization* Many machine learning classifiers have multiple parameters, which significantly affect the performance and behavior. It would be interesting to see, whether the insights of the weight of features could be used to create methods to efficiently optimize the used parameters.
4. *Detection and avoidance of overfitting* As shown in Table 9 classification tasks, especially on small data-sets, tend to result in overfitted classifiers. Using the information on the weight and influence of particular words and phrases could be valuable to detect and also to avoid overfitting of classifiers.
5. *Domain portability* A huge challenge is still the domain portability of machine learning classifiers. If a classifier is trained on one domain, e.g. tenancy law, it might not be used in another domain, such as tax law. However, if the internal structure is more accessible using local-linear explanations the portability might become easier or those instances in the new domain, which could be classified can be determined more easily and the training could be more efficient for the remaining documents.

These considerations show the great potential of local-linear explanations as an additional measure in machine learning for legal text classification. Having machine learning systems that explain their behavior is attractive to domain experts, as they get more efficient and reliable systems, and to engineers who will better understand the behaviour and can improve them even further.

6 Conclusion

This paper describes two experiments on the automated classification of legal norms with regard to their semantic type. Based on legal and linguistic aspects the taxonomy of nine different semantic types was created. It focuses on functional aspects of norms, e.g. duty, permission, prohibition, etc.

We conducted two experiments to exploit the potential of the automated classification: (1) a rule-based approach using hand-crafted pattern definitions, and (2) a machine-learning based approach comparing different classifiers and parameter settings. Based on the German tenancy law, i.e., civil law, we manually labeled 601 norms which were used for training and evaluation.

In the experiments a F_1 score of 0.78 (rule-based) and 0.83 (machine-learning based) for the classification task was achieved. To inspect the machine-learning based approaches a new technique, local-linear interpretable model-agnostic explanations (LIME), was used to make the trained classifier transparent. We showed that the classifier decisions are highly related to the knowledge engineering approach as similar tokens are highly relevant during the classification task. LIME provides additional evidence, to determine overfitting, especially on small datasets and types with low support.

In the digital age, we expect more and more tasks, currently performed by domain-experts, to be done by intelligent and smart systems. We consider this work as a contribution for tasks related to eDiscovery, forensics, and technology assisted review, which is particularly relevant for the legal domain.

References

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

Biagioli C, Francesconi E, Passerini A, Montemagni S, Soria C (2005) Automatic semantics extraction in law documents. In: Proceedings of the international conference on artificial intelligence and law, pp 133–140

Bundesministerium der Justiz (2008) Handbuch der Rechtsförmlichkeit. Bundesministerium der Justiz, Berlin

Chiticariu L, Li Y, Reiss FR (2013) Rule-based information extraction is dead! long live rule-based information extraction systems! In: EMNLP, October, pp 827–832

Cunningham H, Maynard D, Tablan V (2000) Jape: a java annotation patterns engine. Technical report. University of Sheffield, Sheffield

Francesconi E, Passerini A (2007) Automatic classification of provisions in legislative texts. Artif Intell Law 15:1–17

Hart HLA, Green L (2012) The concept of law. Oxford University Press, Oxford

Hohfeld WN (1917) Fundamental legal conceptions as applied in judicial reasoning. Yale Law J 26(8):710–770

Klügl P (2014) Context-specific consistencies in information extraction. Ph.D. thesis

Larenz K, Canaris C (2013) Methodenlehre der Rechtswissenschaft. Springer-Lehrbuch. Springer, Berlin. https://books.google.de/books?id=DeuHBwAAQBAJ

Maat E, Winkels R (2007) Categorisation of norms. In: Jurix: conference on legal knowledge and information systems, pp 79–88

Maat E, Krabben K, Winkels R (2010) Machine learning versus knowledge based classification of legal texts. In: Jurix: conference on legal knowledge and information systems, pp 87–96

Maat E.d, Winkels R (2010) Automated classification of norms in sources of law. In: Proceedings of workshop on semantic processing of legal texts, pp 170–191

Moreso JJ (2014) Benthams deontic logic. In: Tusseau G (ed) The legal philosophy and influence of Jeremy Bentham. Routledge, Abingdon, pp 83–91

Paulo Aires J, Pinheiro D, Strube de Lima V, Meneguzzi F (2017) Norm conflict identification in contracts. Artif Intell Law 25:1–32

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Ribeiro M.T, Singh S, Guestrin C (2016) "why should I trust you?": Explaining the predictions of any classifier. CoRR arXiv:1602.04938

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, August 13–17, 2016, pp 1135–1144

Šavelka J, Ashley KD (2016) Using conditional random fields to detect different functional types of content in decisions of united states courts with example application to sentence boundary detection. Second Workshop on automated detection, extraction and Analysis of Semantic Information in Legal Texts (ASAIL)

Susskind R (2013) Tomorrow's lawyers: an introduction to your future. Oxford University Press, Oxford

Veith C, Bandlow M, Harnisch M, Wenzler H, Hartung M, Hartung D (2016) How legal technology will change the business of law. Technical report. Boston Consulting Group, Boston

Waltl B, Matthes F, Waltl T, Grass T (2016) Lexia: a data science environment for semantic analysis of German legal texts. Jusletter IT 4:4–1

Waltl B, Vogl R (2018) Explainable artificial intelligence—the new frontier in legal informatics. Jusletter IT 4:1–10

Wyner A, Peters W (2010) On rule extraction from regulations. In: Jurix: conference on legal knowledge and information systems