Artificial Intelligence and Law
https://doi.org/10.1007/s10506-024-09415-9

ORIGINAL RESEARCH

Classifying proportionality - identification of a legal argument

Kilian Lüders[1] · **Bent Stohlmann**[1]

Accepted: 19 July 2024

Abstract
Proportionality is a central and globally spread argumentation technique in public law. This article provides a conceptual introduction to proportionality and argues that such a domain-specific form of argumentation is particularly interesting for argument mining. As a major contribution of this article, we share a new dataset for which proportionality has been annotated. The dataset consists of 300 German Federal Constitutional Court decisions annotated at the sentence level (54,929 sentences). In addition to separating textual parts, a fine-grained system of proportionality categories was used. Finally, we used these data for a classification task. We built classifiers that predict whether or not proportionality is invoked in a sentence. We employed several models, including neural and deep learning models and transformers. A BERT-BiLSTM-CRF model performed best.

Keywords Proportionality · Constitutional court · Legal arguments · Transformer · BERT · BiLSTM

1 Introduction

Courts use specific argumentation techniques to justify their decisions. One technique that has received much attention in the public law field is proportionality. Proportionality is constitutional law's most important measure of weighing up the interests of the parties of a proceeding. Beginning with the German Federal Constitutional Court (GFCC), it has spread globally. It is used in the jurisprudence of the European Court of Justice, the European Court of Human Rights, and many constitutional courts, such as Israel, Canada, and South Africa (Steiner et al. 2020; Petersen 2017).

This article seeks to identify proportionality as a specific argumentation technique in GFCC jurisprudence. It, therefore, proposes proportionality as an object

✉ Kilian Lüders
kilian.lueders@hu-berlin.de

[1] Faculty of Law, Humboldt-Universität zu Berlin, Berlin, Germany

Published online: 03 August 2024

for argument mining. An argument is understood *as a sub-unit of a text asserted to offer support to a claim* (Brewer 2018, p. 152–154). Argument mining aims to extract structured argument data from unstructured text to answer questions about *why* something is the case (Mochales and Moens 2009; Lawrence and Reed 2019). Our paper argues that it is worth considering domain-specific argument types for argument mining. We thus turn to the techniques that courts use in their reasoning. As one such argumentation technique, we introduce proportionality. As a working definition, we understand proportionality as *the assessment of value decisions by public authority along the lines of an often multi-step means-end relation*.

To perform argument mining for proportionality, a suitable dataset is required. With this paper, we provide such a dataset to the research community. This dataset contains 300 GFCC decisions annotated at sentence level (54,929 sentences). It includes annotations for the structure of the text and very fine-grained variables for proportionality. We extensively describe the creation process, the inter-annotator agreement, and the expert curation of the gold standard.

Finally, a classification task is carried out on this data set. The task is to classify whether a sentence contains proportionality or not. Different types of models are used: baseline models are tested against more complex neural and deep learning models. We also use variants of the transformer model BERT. A BERT-BiLSTM-CRF achieved the best result in our experiment. Although the model's predictions have a high accuracy (0.905), it struggles to find all sentences with proportionality. No model in the experiment has a recall higher than 0.5. A subsequent error analysis sheds light on the challenges of the task.

This article makes, therefore, three contributions: *First*, we conceptualize proportionality as a legal argument. *Second*, we provide the new dataset. *Third*, we build a corresponding classifier using different neural and deep learning models. The remainder of the paper is organized as follows: In the next paragraph, we will briefly discuss related works (Sect 2). We will then introduce proportionality as a particularly relevant argument technique the GFCC uses (Sect 3). Afterwards, we will introduce the newly created dataset (Sect 4) and the methodological approaches used (Sect 5). Finally, we will discuss the performances of the classifiers (Sect 6).

2 Related Work

As a first step, we briefly highlight related work and provide an overview of current research. We consider the literature on legal argument mining (Sect 2.1), annotation studies on legal case documents (Sect 2.2), previous empirical studies on proportionality in the GFCC's case law (Sect 2.3), and deep learning methods on legal texts (Sect 2.4).

2.1 Legal argument mining

Within the broader literature on argument mining, a subset of studies focuses on extracting arguments from legal case documents (Lawrence and Reed 2019, p. 782; Cabrio and Villata 2018, p. 5430; Lippi and Torroni 2016). Previous studies have mined arguments in documents from various jurisdictions, most notably the European Court of Human Rights. These studies have primarily focused on identifying logical subcategories of arguments, dividing them into premises and conclusions (Mochales and Moens 2009; 2011). Recently, studies have attempted to operationalize specifically legal argumentation structures (Habernal et al. 2023; Gretok et al. 2020; Bhattacharya et al. 2019; 2023; Grabmair et al. 2015). Grabmair et al. (2015) classify various legal documents related to vaccine injury compensations in the US using a label for "legal rules" and one for "evidence-based findings:" (Grabmair et al. 2015, p. 71). Bhattacharya et al. (2019, 2023) identify seven "rhetorical roles" of text units in Indian and UK court decisions. They classify roles like "argument", "facts" or "ruling by present court" (Bhattacharya et al. 2023, p. 61). Gretok et al. try to classify which of two types of rules US courts have applied in Fourth Amendment cases (protection from unreasonable searches and seizures by the state). They distinguish between "bright-line" and "totality of the circumstances" rules, taking the concepts from legal literature (Gretok et al. 2020, p. 63). Most related to our study, Habernal et al. (2023) classify decisions of the European Court of Human Rights according to numerous categories devised collaboratively with legal experts. The categories are structured around the three meta-categories "procedural arguments", "method of interpretation", and "test of the principle of proportionality" (Habernal et al. 2023). While there is prior work on the semantic classification of German legal documents (Waltl et al. 2019), no automated legal argument mining exists for German court decisions in general and those of the GFCC in particular.

2.2 Annotation studies on legal case documents

As legal concepts are often ambiguous and overlapping, they are hard to identify, even for human annotators. As high-quality datasets are essential for any ML project, reporting on the annotation process and the quality of the annotation data is increasingly important. However, only some studies give detailed information on their annotation process or data quality (Lawrence and Reed 2019, p. 806; Cabrio and Villata 2018, p. 5432). Wyner et al. (2010) provide an in-depth study of annotation processes and curation of gold standards for machine learning in the field. Correia et al. (2022) develop a corpus of 595 excerpts from Brazilian Supreme Court decisions and provide a comprehensive report on the creation process and agreement measures. Shulayeva et al. (2017) extensively describe their annotation process and briefly report on the inter-annotator agreement for 50 UK law reports. Bhattacharya et al. (2023) also report various agreement measures for a complex annotation task for a gold standard dataset of 100 case documents from UK and Indian case law.

Existing studies rely on the majority decision to merge annotations into a gold standard (Correia et al. 2022, p. 12; Bhattacharya et al. 2023, p. 66). Only Habernal et al. (2023) employ independent expert annotators to solve disagreements and curate their gold standard. No prior work exists that does such thorough reporting on a gold standard data set of the scale offered here (300 case documents).

2.3 Quantitative approaches to proportionality

There has been previous work on identifying and assessing proportionality in the GFCC's case law. Petersen (2017) investigates the use of proportionality as an argument in the striking down of laws in the case law of the GFCC and the case law of the Canadian and South African constitutional courts. He, therefore, manually annotated for a set of 250 decisions of the GFCC whether the decisions invoke different argument types (including proportionality) to strike down laws. Lang (2020) investigated whether a subset of 114 GFCC decisions invoke proportionality. Neither of these studies annotates at the sentence level, cross-annotates with multiple annotators, or provides much information about their annotation process and its assessed quality. They do not automate any classification process or publish any annotated corpora.

2.4 Deep learning with legal text

Deep learning on legal text has a wide range of applications, including legal named entity recognition (Correia et al. 2022; Leitner et al. 2019), information and document retrieval (Nguyen et al. 2022; Zhu et al. 2023), decision outcome prediction (Valvoda et al. 2023; Mumcuoğlu et al. 2021), summarization (Zhong et al. 2019), as well as a variety of other classification tasks (Costa et al. 2023; Cohen et al. 2023).

Bhattacharya et al. (2023, 2019) have explored various deep-learning methods. They use BiLSTM architectures to classify the entire sequence of sentences in court decisions. Their experiments range from BiLSTMs with newly trained word embeddings to models that combine BERT transformers with BiLSTMs and CRFs. In particular, the latter models performed well, as did BiLSTM models using domain-specific pre-trained embeddings. Of note is the study by Habernal et al. (2023), who, relying on transformer models, employ various pre-training and fine-tuning strategies to identify arguments in ECHR decisions. Thalken et al. (2023) recently published a paper distinguishing between different types of arguments in US Supreme Court decisions, comparing the performance of transformers with several large language models. They find that trained transformers deliver significantly better classification results.

Overall, transformers such as BERT provide significant momentum to the research field (Greco and Tagarelli 2023; Devlin et al. 2018). There is also a specific legal BERT for the English language (Chalkidis et al. 2020). Several studies have shown that using domain-specific BERT variants leads to strong results (Chalkidis et al. 2019; Habernal et al. 2023; Thalken et al. 2023). However, research to date has mainly focused on English legal texts. There is less experience and a lack of large and well-trained models for other languages.

3 Proportionality as a legal argument

In the following, we will look at the conceptual aspects of our project. It is worth considering domain-specific argument types for legal argument mining (Sect 3.1). We present proportionality as such a domain-specific argument type (3.2.), and discuss our conceptual considerations (Sect 3.3).

3.1 Argument mining and domain-specific arguments

The application of argument mining to judicial reasoning requires some preliminary conceptual thoughts. What can be understood as an argument in the legal context is a complex question.

The concept of argument can be defined *as a sub-unit of a text asserted to offer support to a claim* (Brewer 2018, p. 152–154). Following the basic categories of logic, one can distinguish between premises and claims in an argument. The next step is to consider how this understanding of arguments relates to court decisions (Toulmin 2003, p. 7). Judges explicate their reasoning in judicial decisions. However, when they do so, they do not structure it according to the categories of logic. Arguments are often part of a nested, interleaving structure. One argument's claim is another argument's premise in a cascade of arguments leading to the decision's outcome.

Furthermore, courts create their own argumentation structures and layouts. One could even say more theoretically that law has its own criteria to structure reasoning. Courts also follow these criteria in giving their reasons (Möllers 2013, p. 89f.). Theory of argument since Aristotle accounted for this possibility of arguments whose logical form is not fully explicated in practice (Rapp 2023). The concept of *enthymeme* in Aristotle's Rhetoric allows for arguments which do not explicitly state all their premises. Brewer also includes under this concept all arguments "whose mode of logical inference is not explicit in their original mode of presentation" (Brewer 2018, p. 155). He goes on to note that arguments offered by judges are overwhelmingly *enthymematic*.

Argument mining tries to turn unstructured text into structured data to understand *why* something is the case (Mochales and Moens 2011, p. 1f.; Lawrence and Reed 2019, p. 806). To this end, many studies attempt to automatically distinguish premises from conclusions (e.g. Mochales and Moens 2011). Considering the subordinate role of these categories in structuring judicial reasoning, a more promising starting point to fulfill this task is to look at domain-specific arguments. Then, the task turns to the differentiation between types of arguments. Again, argument mining is concerned with answering why questions. Thus, it has to consider the requirements and structures that guide the arguer in her specific context. If we want to find the answer to why a court made a decision in its reasoning, we have to look at the requirements of law.

Moreover, it shows a way forward in operationalizing arguments for creating high-quality annotated legal text corpora. Therefore, the way to go is to involve experts in the field of judicial argumentation. At the moment, there is still room for

improvement in this area of research. As Habernal et al. (2023) correctly pointed out, there is a significant gap between what is understood as an argument in NLP and what a legal perspective on judicial reasoning looks like.

3.2 Proportionality as a cornerstone of constitutional adjudication

Proportionality is a constitutional argument that *assesses a value decision by public authority along the lines of an often multi-step means-end relation*. In the structure of a judicial decision, it is a test for the justification of the infringement of an individual right through the public authority's measure. Often, the constitutionality of a public measure depends on the court's assessment of its proportionality. In constitutional law, proportionality is considered the most dominant argument type (Stone Sweet and Mathews 2019, p. 4f.; Steiner et al. 2022, p. 642f.). It is even regarded as "omnipresent" (Peters 2021, p. 1138; Möller 2015, p. 13) and a basic criterion of legal validity (Stone Sweet and Mathews 2019, p. 9; Möllers 2020, p. 163). Moreover, it is attributed with a reshaping effect on the separation of powers and the political system's governance structure (see again Stone Sweet and Mathews 2019, p. 161, 127ff.).

Proportionality is constitutional law's most important measure of weighing the parties' interests in a proceeding. It serves as a central measure to assess and justify value decisions. This centrality to the justifications of constitutional courts in general and the GFCC in particular (see e.g. Tischbirek 2020, p. 14ff.; Lepsius 2020, p. 113) alone would sufficiently demonstrate the relevance of identifying proportionality arguments in the GFCC's decisions. Internationally, proportionality as an analytical framework deployed by constitutional courts is also considered one of the most relevant examples of the migration of constitutional ideas (Weinrib 2007, p. 84ff.). The practice of the German constitutional court is often considered exemplary from a genealogical and structural perspective (Stone Sweet and Mathews 2019, p. 60ff.; Lepsius 2020, p. 95; Peters 2021, p. 1136).

Definition Surprisingly, there is little effort in legal discourse to give an actual definition of proportionality. It can be described as a structured balancing test consisting of a means-ends comparison comprising multiple steps (Huscroft et al. 2014, p. 2; Grimm 2007, p. 387; Stone Sweet and Mathews 2019, p. 35). For the GFCC, the proportionality argument is often described as consisting of four steps, namely the existence of a "legitimate aim", the "suitability", "necessity" and "balancing" (cf. Steiner et al. 2022, p. 647f.; Peters 2021, p. 1136f.; Petersen 2017, p. 80). Therefore, in scholarly practice, proportionality is seen as a considerably formalized type of constitutional argument (Huscroft et al. 2014, p. 2). This makes proportionality the perfect object for a legal argument mining study.

Last but not least, proportionality corresponds to the goal of argument mining to answer a why question adequately: Why, for example, did the GFCC declare a government measure unconstitutional? Because it was disproportional insofar as the

means were not proportional to the measure's end. As a working definition, *proportionality is the assessment of value decisions by public authority along the lines of an often multi-step means-end relation*. It must be noted that this definition does not allow for a comprehensive demarcation of the concept.

Example Table 1 gives an example of a proportionality test. It is an excerpt of a decision included in the dataset (BVerfGE (official collection of GFCC decisions) Volume 68, page 272). The decision is concerned with particular passages of the Hessian state building code. They require certain educational qualifications for the right to approve documentation needed for planned construction. As set out above, proportionality tests the justification of a public authority measure. Here, the court already determined that the relevant section of state construction law interferes with the applicant's freedom to practice an occupation. It is now concerned with the possible justification of such an interference.

Sentences 1–3 consider the reasons the legislator gave for implementing the measure. It amounts to assessing the legitimate aim pursued through the measure, the first step of the proportionality test. Sentence 4 then states the suitability of the measure to pursue this aim, thus accounting for the second step of proportionality. Moreover, it determines the measure to be "necessary", the third proportionality requirement. The following sentence (Sec 5) then further elaborates on why this measure is considered to be necessary. Finally, sentences 6–10 balance the objectives of the measure and the rights interfered with. Taking into account the regulation's differentiations, the Court assesses in detail whether such a measure is justified. This accounts for the final step of the proportionality test. Overall, sentences 1–10 thus are considered one test of proportionality.

In this example, the proportionality test spans a total of ten sentences and consists of all four steps in their typical order. However, the proportionality test can deviate substantially from the more prototypical structure of this example. Sometimes, the test can be interrupted by passages of text that do not themselves amount to parts of the proportionality test. One decision can include multiple passages containing different proportionality tests of other public authority measures and additional rights the measure interferes with. The test does not have to give all of the customary four steps and sometimes is not distinguishable into steps at all. As also seen in the example, the court sometimes explicitly names the steps in its reasoning (see "suitable" and "necessary" in Sentence 4). Other times, the steps are not explicitly addressed under their familiar names (see legitimate aim and balancing steps in sentences 1–3 and 6–10). The length of the test can vary from only one sentence (often just stating a measure's compliance with the test) to 84 paragraphs (314 sentences) or even more.[1] These examples illustrate that the GFCC uses this type of argumentation very heterogeneously in practice and that the classification task is, therefore, not trivial.

[1] Cf. GFCC: BVerfGE 150, 244. Order of the First Senate of 18 December 2018. (English Translation). http://www.bverfg.de/e/rs20181218_1bvr014215en.html Accessed 9 June 2023.

Table 1 Example of a proportionality test. Selected paragraphs of the GFCC decision BVerfGE 68, 272

Sentence	GFCC decision	Label
1	In the original draft bill, the demand for a specific qualification of the author of the design was based on the fact that this was intended to achieve something in the public interest that could not be achieved by building supervision, namely a general improvement of the quality of construction with regard to economic efficiency, rational design, and functionality of the buildings, but not least also with regard to building culture (LTDrucks. 8/55 S. 108 f.)	Aim
2	This reasoning does not preclude the introduction of a special building permit authorization from being justified on the grounds of public safety	Aim
3	Even if the building permit authorities are obliged to reject building submissions that have been designed contrary to the rules of architecture, on the basis of incorrect structural calculations or in disregard of building regulations, the legislator can, in the interest of increased safety and also to relieve the building permit procedure, require that the necessary submissions are already prepared and are the responsibility of experts with the appropriate training and experience (cf. BverfGE 28, 364 [375]; BayVGH {Bavarian Higher Administrative Court – translator's note}, BayVBl. 1978, p. 207 [209]; see also Rasch/Schaetzell, Hessische Bauordnung, in: Die Praxis der Gemeindeverwaltung, F 3 He, p. 293 f.)	Aim
4	b) The means chosen by the legislator to require the person authorized to submit plans to acquire a specific professional qualification was suitable and necessary to achieve the legislative objectives	Suitability Necessity
5	In particular, it is not apparent how these objectives could have been achieved by other regulations that had less impact on the exercise of the profession	Necessity
6	Furthermore, the minimum qualification required for the authorization to submit plans for simpler construction projects is not unreasonably high, but rather enables not only architects and construction engineers but also master craftsmen and their equals to recognize construction plans to a considerable extent within the scope of their craft activities	Balancing
7	At most, it could be doubtful, whether all of the simpler building projects mentioned in § 91 s. 4 Hessian building code are not only to be treated as requiring authorization but must also be the responsibility of a design author authorized to submit plans	Balancing
8	However, the legislator has already differentiated between projects requiring approval, for which only architects and construction engineers are authorized to submit plans, other simpler building projects requiring approval, for which master craftsmen and equivalent persons may also approve building plans, projects requiring notification only (§ 88 Hessian building code) and finally projects not requiring approval or notification (§ 89 Hessian building code and the exemption ordinance)	Balancing
9	A further differentiation cannot be imposed on the legislator by the constitution	Balancing
10	There are all the less constitutional objections to the more detailed differentiation incumbent on the legislator as the building supervisory authority can dispense with the appointment of a design author in individual cases for technically simple structures (§ 77 s. 3 Hessian building code)	Balancing

Translated by the Authors. The sentence count in the left column is information added by the authors and does not reflect the text structure of the original document. The original german version can be found in Appendix A

3.3 Conceptual considerations

Before we turn to our dataset, two conceptual aspects need to be clarified:

First, we would like to emphasize that proportionality can only occur in the merits of a decision. We have also emphasized this point in the annotation guidelines. The other parts of the text, particularly the facts of the case and admissibility, are not the subject of this study. This critical but not common step of limiting the data ensured that only text passages in which the court was concerned with giving its arguments were included in our dataset. Otherwise, there is a risk that the data would include extensive passages in which the court repeated the parties' arguments or prior stages of appeal, possibly polluting the data.

Second, it is noteworthy that we are interested in the court's invocation of proportionality. Most attributions of relevance to proportionality mentioned above imply the need for proportionality to be the *deciding* argument for the outcome of certain decisions. However, the court often does not indicate which argument was decisive. We do not differentiate between proportionality's mere invocation and its application to the facts of a case. Instead, we focus on identifying whether or not the GFCC is invoking proportionality as an argument.

4 Data

In the following, we will introduce our newly annotated dataset. A set of 300 GFCC decisions was annotated at the sentence level. The published dataset contains many resources, including the individual annotations, the curated gold standard, and the guidelines (Lüders et al. 2024). It can be used for many projects beyond this publication, including other argument mining projects. But it also includes annotations on the textual parts of decisions,[2] for example.

The dataset was created in collaboration with lawyers pursuing empirical-descriptive research interests.[3] As a result, the dataset contains very fine-grained annotation categories, some of which occur only a few times. This paper is the first study to use the data for a classification task.

4.1 Data selection and annotation

The data basis for our study is a collection of texts from all decisions of the GFCC, which includes over 10.000 documents (Wendel and Möllers 2023). Our research focuses on the decisions of the court's two senates. Therefore, the base population consists of 3.371 decisions published by the two senates in the court's official collection of decisions, starting with the first decision in 1951 to 2021. For our study,

[2] This type of data may be particularly interesting for rhetorical role labeling projects (e.g., Saravanan et al. 2008; Šavelka and Ashley 2016; Bhattacharya et al. 2023).

[3] This interest resulted in a descriptive-empirical article in a German law journal, which also uses the data set: Stohlmann et al. 2024.

Table 2 Number of sentences in different text sections of the decisions in the dataset

Textual parts of the decisions	Number of sentences
Facts of the case	25.050
Admissibility	4.717
Overlap (admissibility and merits)	127
Merits	24.577
Signatures of the judges	458

we randomly drew 300 decisions from this set,[4] annotated at the sentence level. It should be noted that GFCC decisions can be very long. The 300 selected decisions fill 5.739 printed pages in the court's official collection.

The annotation's first step was separating the decisions' textual parts. A GFCC decision consists of the sections: "facts of the case", "admissibility", and "merits", as well as the "signatures of the judges". In exceptional cases, there may also be an "overlap" between admissibility and merits. All 54.929 sentences of the dataset were annotated according to these categories. Table 2 contains information on the frequency of the individual categories.

It is important to emphasize that not all decisions comprise all four text sections: 55 decisions have no merit section (for more information, see Table 5 with key statistics of the dataset). This is because sometimes only the admissibility of a case is decided.

The text section variable may be interesting in itself; here, it is an essential prerequisite for further annotation: Proportionality can only be found in the merits section. Accordingly, the proportionality annotations are only applied to a subset of the dataset. All sentences that were not part of the merits were excluded. As a result, there are fewer sentences to annotate for proportionality (only 24.577 sentences). There are also fewer decisions (only 245 decisions) because not all decisions have a merits section. These merits passages were then annotated according to detailed legal categories.

As mentioned above, the annotation process relied on the working definition of proportionality as a *multi-step* means-end relation. Thus, our annotators were asked to annotate for each sentence whether or not it refers to one of the customary four steps, "legitimate aim", "suitability, "necessity" and "balancing". The four-step categories can occur in combination (i.e., they are not disjoint). In the example above (Table 1), Sentence 4 would be labeled as part of both "suitability" and "necessity". If a sentence did not fit a specific step but nonetheless was considered to be part of the invocation of proportionality, they were assigned the label "proportionality unspecific". This accounts for the difficulty of operationalizing a legal concept like

[4] We drew a random sample from all decisions in our base population. We have not weighted, filtered, or manually checked the selection.

Table 3 Inter-annotator agreement (Fleiss ' κ) by annotation cycle and category

Annotation cycle	Aim	Suitability	Necessity	Balancing	Unspecific
Cycle 1	0,085	0,415	0,608	0,600	0,018
Cycle 2	0,403	0,406	0,474	0,817	0,424
Cycle 3	0,667	0,582	0,805	0,712	0,282

Further documentation of the calculation of Inter annotator agreements see Stohlmann et al. 2024

proportionality on a sentence-based level, as they often interleave in the text and are not unambiguously demarcated.

The annotators were given detailed guidelines[5] for this process, which were slightly adjusted between rounds. As this problem applies not only to the concept of proportionality in general but also to the concepts of the four steps we asked annotators to identify, we provided them with similar working definitions to the one of proportionality but stressed that these were to be considered more of a guidance structure than an also unequivocal demarcation. Instead, we relied on the expertise our annotators gained through their legal training to identify the legal concepts. For example, the annotation guidelines state concerning all categories that "the given definitions are supposed to describe an ideal type. However, we also aim to trace the developments of the concept outside of its canonical appearance. Therefore, you should also annotate the proportionality test in instances in which you recognize it due to your legal education, even though it does not fulfill 'classical' requirements."[6]

Besides general remarks, the guidelines for the "necessity" step, for example, only state that necessity should be annotated "if the question of possibly less intrusive measures is addressed". A few ideal-type examples supplement this.[7]

Annotation has taken place in three cycles. It is now considered good practice to have more than one person annotate each document. In the first two cycles, each decision (29 and 56, respectively) was annotated by three annotators, and in the last cycle (215 decisions) by two annotators. In total, 13 persons participated in our annotation effort, all having studied law in Germany for at least four semesters.

4.2 Inter-annotator agreement and gold standard curation

The cross-annotation by multiple annotators makes documenting Inter Annotator Agreements (IAA) necessary. Table 3 shows the IAA for the different annotation categories divided by annotation cycles. Table 4 then shows the IAA for a general

[5] The annotation guidelines can be found in the data repository (Lüders et al. 2024).

[6] The authors translated this and the following quote from the guidelines. The original annotation guidelines in German can be found in the data repository (Lüders et al. 2024).

[7] For each sentence with proportionality, a context variable was annotated concerning whether it is a test regarding civil liberties, equality rights, or the law of state organization. This information is not displayed in the example above as it is not used for this study. It is just reported to give a complete account of the annotation process.

Table 4 Inter-Annotator Agreement (Fleiss ' κ) by annotation cycle for annotation of proportionality

Annotation cycle	Proportionality (collapsed)
Cycle 1	0,560
Cycle 2	0,781
Cycle 3	0,776

proportionality variable. This variable was created by collapsing the other categories into one. If any of the "step-categories" (f.e. suitability) or the category "unspecific" were annotated, the sentence was treated as being annotated as "proportionality". This collapsed variable is also the target variable of our task to classify sentences. IAA was calculated using the Fleiss ' κ measure (cf. Fleiss 1971; Artstein and Poesio 2008). Fleiss ' κ shows the agreement beyond pure coincidence. Its values range from -1 to 1, with 0 being coincidental agreement.

The IAA for the collapsed variable differs between cycles: For the first round, Fleiss ' κ was 0.56, in the second, 0.78, and the last, 0.78. These values show that identifying the proportionality test is not straightforward, even for law students. Compared to other legal argument mining studies, there is a slight drop-off in IAA (cf. Bhattacharya et al. 2023, p. 63). However, Bhattacharya et al. found that agreement differs strongly between different labels. Our study focuses on labeling proportionality and its sub-units. This is a label that is itself challenging and, thus, can be expected to produce lower IAA scores. This matches Bhattacharya et al.'s assumption that different levels of agreement in their study stem from the differentiation between specific labels, which is more challenging from a legal perspective (2023, p. 65). Comparing the different IAA for different categories set out in Table 3, the same seems to hold for our study. While, for example, the balancing step of proportionality seems easier to annotate, the "unspecific"-category shows comparably poor IAA. The balancing step is a relatively well-known concept in law with the precise function of weighing up conflicting interests and values. In contrast, the "unspecific" category includes all forms of proportionality that do not fit the classic scheme. Thus, agreement for these text units is expected to be lower as no or little legal criteria exist.

We had to curate a gold-standard dataset of sentences and their labels to prepare the annotated data for machine-learning tasks. Thus, all cases in which not all annotators agreed were decided by an expert in a separate evaluation. Particularly complicated cases were discussed by a group of experts. Experts in this step were only jurists who had finished law school and had academically worked on proportionality. This extra step aimed to provide additional quality of gold standard data compared to using, for example, majority rules for curation. Considering the disagreement among expert annotators, this extra curation step is precious in obtaining high-quality gold-standard data. The approach resembles the curation process of Habernal et al. (2023). A dataset without annotation dissent was created to be used as a gold standard for ML training.

Table 5 Key statistics on sentences and decisions with proportionality in the dataset

Number of decisions	300
→Decisions with merits section	245
→Thereof with proportionality	60
Number of sentences	54.929
→Number of sentences with merits section	24.577
→Thereof with proportionality	2.870

4.3 Dataset characteristics

For our particular classification task, the multiple annotation categories of our data (see above) were collapsed into a new variable: If at least one of the steps of the proportionality test or the category for unspecific occurrences of proportionality were annotated, the sentence is considered to invoke proportionality. Analogously, if at least one sentence in the decision was annotated, the decision is deemed to contain a proportionality test. Table 5 shows the key statistics on decisions and their sentences with proportionality in the data set.

As can be seen, proportionality is not omnipresent. It appears in only 60 decisions, and only 245 decisions contain a section on the merits. Among the sentences, proportionality occurs in only 11.7% of the merit sentences. This means that the data for the classification task is unbalanced (out of 24.577 sentences, only 2.870 belong to the target group).

As mentioned above, proportionality and, in principle, decisions by the GFCC can vary greatly. Figure 1 illustrates the key characteristics of the data: First, it should be emphasized that GFCC decisions are very heterogeneous in length. Figure 1a) shows a boxplot of the merits' length. The graph shows that the majority of merits are less than 100 sentences long. However, there are quite a few very lengthy outliers. Second, proportionality tests can also vary widely in length. Figure 1b) shows a boxplot for the length of the proportionality tests. The picture is similar: the majority of the proportionality sequences are not longer than 25 sentences, and some are just one sentence long. However, there are, again, extremely long tests. Figure 1c) shows the proportion of the merits in which the proportionality test occurs. Furthermore, very heterogeneous constellations can be observed: In some cases, proportionality accounts only for a small part of the decision's merits. In others, it makes up as much as 80 percent of the decision's merits.

5 Methods

We now turn to our classification experiment. The task is to classify whether or not sentences on the merits of a decision are part of a proportionality test. The target variable is, therefore, the collapsed proportionality variable, as explained

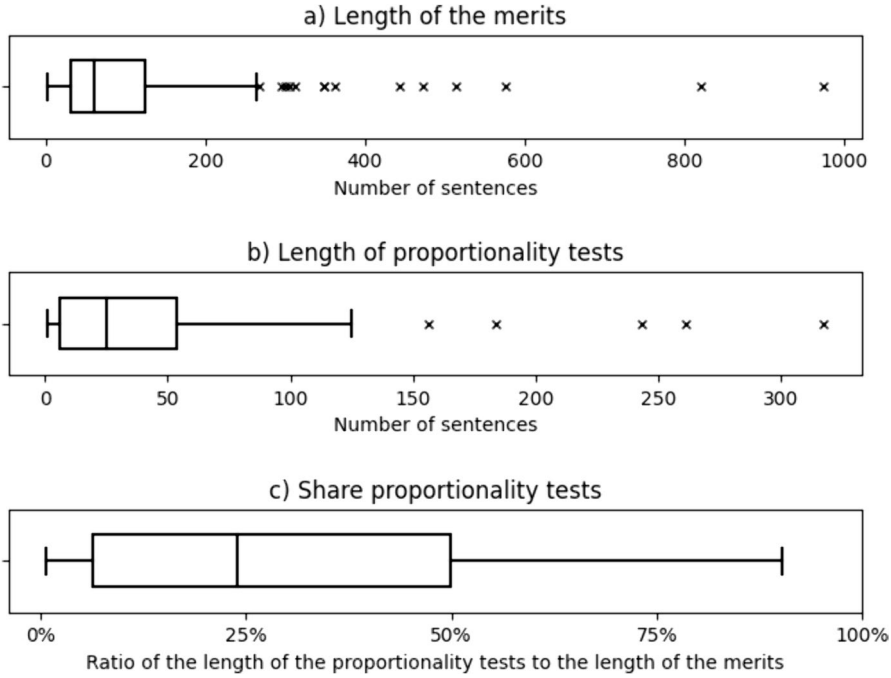

Fig. 1 Descriptive statistics for the annotated proportionality test data set. Boxplot **a** shows the length of the merits (number of sentences for all decisions with merits). Boxplot **b** shows the length of the proportionality tests (number of sentences for all decisions with proportionality). Boxplot **c** visualizes the share of proportionality tests in the merits (for all decisions with proportionality)

above (Sect 4.2). The database of our experiment consists of all sentences of the merits (24.577). This is, therefore, a binary classification problem, whereas the decisions are represented as a sequence of sentences.

Several approaches are used in the experiment: In addition to a majority and a classical ML model, a rule-based approach that recognizes typical phrases of the proportionality test is used as a baseline. All baseline approaches classify the sentences independently. However, we expect it will often be challenging to classify sentences individually. It is frequently the case that only the context indicates whether or not a sentence is part of a proportionality test. This assumption is based on the annotation process and the feedback from the annotators. We use neural networks with bidirectional Long Short-Term Memory (BiLSTM) layers for the classification task (Graves 2012; Graves and Schmidhuber 2005). This technique allows information to be shared across sentences.

We train a range of configurations for the BiLSTM models: We work with completely newly trained word embeddings but also use pre-trained embeddings. Additionally, we also make use of BERT models and combine them with BiLSTM. We are also experimenting with Conditional Random Fields (CRF) on top of the BiLSTM layers in all these setups. The combination of BiLSTM and CRF is common

and widely used for sequence tagging and NER (e.g. Correia et al. 2022). Bhattacharya et al. (2023) and their best-performing models guided our model selection and configuration, as they also worked on a similar task of classifying sentences in legal decisions.

5.1 Baseline

To categorize the performance of more complex models, we create baselines. First, a majority model is introduced. This always predicts the most frequent category, i.e., no proportionality. The model's primary purpose is not to overestimate the performance of models on a dataset where only about 12% of the sentences are coded proportionality (see Table 5).

We use a classical ML approach and a Support Vector Classifier (SVC) as a further baseline. When working with legal text data, SVCs receive much attention because of their excellent performance (Clavie and Alphonsus 2021, p. 4). We used the SVC with two types of features: On the one hand, it is fitted to tfidf vectors (a classical bag-of-words strategy). Therefore, we use consistent tokenization throughout the project based on Spacy.[8] On the other hand, it is fitted to the pre-trained sentence embeddings. We used a word embedding model trained exclusively on GFCC decisions, upon which we averaged sentence embeddings (see Appendix B for details).

There is a whole set of phrases and keywords that are typical for proportionality. Previous work has used these to identify decisions with a proportionality test (Lang 2020). Against this background, we also include a rule-based approach in our experiment. The rules were created by legal experts using their knowledge from the annotation. During the annotation process, extensive experience was gained with the wording used by the Court when talking about the proportionality test. In addition to apparent keywords such as "proportionality", linguistic features such as word types and morphology are used for the classification. The exact mechanism of the rule-based approach is documented online.[9]

5.2 Embedding BiLSTM models

Four different BiLSTM models using embeddings were trained. The model architectures were proposed by Bhattacharya et al. (2019) and successfully implemented. Since they were already able to outperform baseline models, we have reused them.[10] Whether they are also feasible for German texts and argumentation data remains to be tested. The models are visualized in Fig. 2 and explained below.

[8] Of course, BERT comes with its own tokenizer. However, in all other scenarios, we use the Spacy model *de_core_news_lg*. Available at: https://spacy.io/models/de

[9] https://github.com/klueders/prop_rules

[10] Resources of Bhattacharya et al. (2019) are available at: https://github.com/Law-AI/semantic-segmentation/tree/master

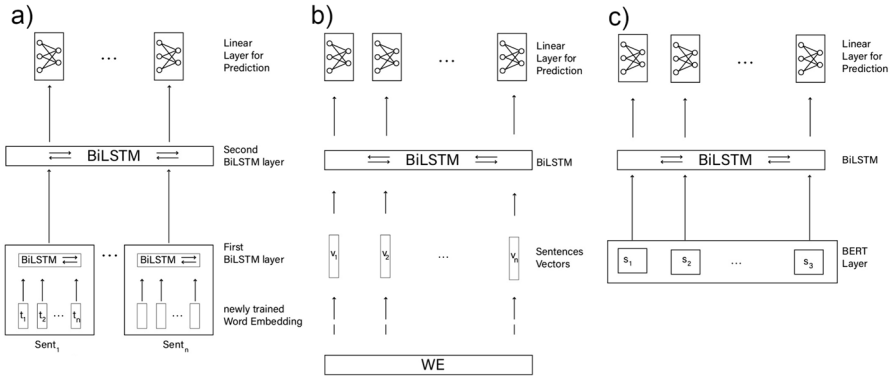

Fig. 2 Graphic visualizing the architecture of the model. **a** BiLSTM with newly trained embeddings; **b** BiLSTM with sentence embeddings; **c** BERT-BiLSTM. The CRF versions of the models are not shown

5.2.1 Newly trained embeddings with BiLSTM

The first model type trains new word embeddings from scratch and consists of two layers of BiLSTM (shown in Fig. 2a). The first BiLSTM layer takes the randomly initialized word embeddings as input and creates 200-dimensional embeddings for each sentence. These sentence embeddings are the input to the second (higher) BiLSTM layer. This considers the whole sequence of sentences and creates another 200-dimensional hidden vector for each sentence. A linear layer finally uses this output vector from the second layer to predict each sentence. The strategy behind this type of model is to train a model solely from the textual data without using any additional information.

5.2.2 Sentence embeddings with biLSTM

The second type of model uses pre-trained sentence embeddings (shown in Fig. 2b). As described above, each sentence is represented as a 200-dimensional vector. Accordingly, the model requires only one BiLSTM layer. (The purpose of the bottom layer of the first model has already been fulfilled by the sentence embeddings). The remaining BiLSTM layer then takes the vectors of all the sentences of a decision as input and produces a 200-dimensional hidden vector for each sentence. Again, a linear layer is applied to predict each sentence. The strategy behind this model is to incorporate additional knowledge about the word usage of the GFCC.

5.2.3 Adding a CRF

In addition, both models just presented were re-created, this time with a conditional random field. They are entirely analogous to the ones described above. The only difference is that a conditional random field is added to the last linear layer. Accordingly, the two more models are a newly trained embedding BiLSTM-CRF and a

pre-trained sentence embedding BiLSTM-CRF. The use of BiLSTM and CRF is a common technique to improve performance and has been applied to legal data (Correia et al. 2022; Leitner et al. 2019).

5.2.4 Implementation details

The models were built using Pytorch (Paszke et al. 2019). For the CRF layer, a ready-to-use and widely used implementation of a linear chain CRF was used.[11] All models were trained unbatched with 200 epochs. An AdamW optimizer with a learning rate of 5e-5 was used. For the models without CRF, a cross-entropy loss was calculated.

5.3 BERT-based models

In addition to the previous models, approaches based on BERT are also implemented. BERT is a transformer model that has become very popular (Devlin et al. 2019). It has also been successfully used in legal argument mining (Habernal et al. 2023; Thalken et al. 2023; Bhattacharya et al. 2023). Unfortunately, there is no German legal BERT, so we used two BERT models trained on general language. On the one hand, we use the original *bert-base-multilingual-uncased* (Devlin et al. 2019),[12] which is often used for German-language classification tasks. On the other hand, we use the *dbmdz/bert-base-german-uncased*, a genuine German BERT model created by the Bavarian State Library.[13] We will use both BERT variants in different scenarios:

5.3.1 BERT with a linear layer

First, we will add only a linear layer to the selected BERT models in a baseline setting. The BERT model creates a 768-dimensional vector for each dataset. The additional linear layer uses this vector to make a prediction. It is, therefore, a baseline sentence classifier that classifies each sentence individually (independently of others). On this basis, we can assess to what extent the use of BERT alone makes a difference.

5.3.2 BERT with BiLSTM

BERT is also used as an input for a BiLSTM layer (shown in Fig. 2c). BERT turns each sentence into a 768-dimensional vector. The sequence of all sentence representations is then captured by a BiLSTM, which returns a 200-dimensional hidden layer for each sentence, as in the models above. A linear layer outputs a prediction on this

[11] Available at: https://github.com/kmkurn/pytorch-crf#egg=pytorch_crf
[12] Available at: https://huggingface.co/bert-base-multilingual-uncased
[13] Available at: https://huggingface.co/dbmdz/bert-base-german-uncased

basis. Again, there are two variants of this model: In one case, the output of the linear layer is used directly as a prediction. In the other case, a CRF is added.

5.3.3 Implementation details

In addition to Pytorch and the CRF layer, the HuggingFace transformers library (Wolf et al. 2020) was used for implementation. The models were again trained unbatched with an AdamW optimizer, but this time with a learning rate of 5e-6 in 50 epochs.

5.4 Performance evaluation

We use a stratified sampled fivefold cross-validation strategy with 80:20 splits. The folds are created so that the proportion of decisions with and without proportionality is retained in both the test and training sets. All models are tested on identical folds. We use scikit-learn (Pedregosa et al. 2011) to create the folds.

Our task is a dichotomous classification problem, and we are interested in one of the categories. Accordingly, we consider.

- **True Positive** (TP): proportionality correctly predicted
- **True Negative** (TN): no proportionality correctly predicted
- **False Positive** (FP): proportionality incorrectly predicted
- **False Negative** (FN): no proportionality incorrectly predicted

On this basis, we use the typical metrics for binary classification problems (Powers 2011):

- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **F1 score** = 2 * (precision * recall) / (precision + recall)

Finally, we also report the **Accuracy** = (TP + TN) / (TP + TN + FP + FN). This is the proportion of correct predictions out of all predictions. An average is calculated for all four metrics over the five splits.

6 Results

In the following section, we will first present and discuss the performance of our model (Sect 6.1.). This is followed by a more detailed error analysis, which takes a closer look at the problems of the best-performing model (Sect 6.2).

Table 6 Classification Performance

Model	Feature	Precision	Recall	F1-Score	Accuracy
BERT-BiLSTM-CRF	Bert-multilingual	0,696	0,343	0,448	0,897
BERT-BiLSTM-CRF	Bert-german	0,708	0,478	**0,561**	**0,905**
BERT-BiLSTM	Bert-multilingual	0,625	0,434	0,483	0,889
BERT-BiLSTM	Bert-german	0,782	0,410	0,491	**0,905**
BERT	Bert-multilingual	0,445	0,282	0,344	0,863
BERT	Bert-german	0,486	0,327	0,390	0,870
BiLSTM-CRF	New embedding	**0,963**	0,032	0,058	0,875
BiLSTM-CRF	Sentence embedding	0,519	**0,499**	0,489	0,865
BiLSTM	New embedding	0,863	0,024	0,044	0,874
BiLSTM	Sentence embedding	0,556	0,446	0,488	0,880
Rule Based		0,893	0,111	0,197	0,884
SVC	Tfidf	0,824	0,072	0,132	0,879
SVC	Sentence embedding	0,724	0,147	0,241	0,884
Majority		0,000	0,000	0,000	0,872

Results of fivefold stratified shuffled cross-validation; reported are means

6.1 Performance

The performance results are shown in Table 6. The Accuracy column shows the proportion of correct predictions made by the models. The values are generally relatively high: even the worst model (BERT with a simple linear layer based on *bert-multilingual*) has an average of 86% correct predictions. The fact that proportionality is rare plays a crucial role here. This is reflected in the performance of the majority model, which never predicts proportionality but still correctly classifies an average of 87% of the sentences. Therefore, when discussing the performance of the models, the F1 score is of particular interest. This score gives insight into how well the individual models can recognize sentences with proportionality.

The baseline models (Majority, SVCs, and Rule-Based) appear not very convincing overall in identifying sentences with proportionality: the F1 scores are all below 0.25. The SVC achieved the best result with pre-trained sentence embeddings (F1 score = 0.241). It is worth noting that the rule-based approach has a comparatively high precision (= 0.893). This means that the sentences identified by this method as having proportionality are very likely to be also sentences with proportionality. At the same time, however, the method is not very suitable for identifying sentences with proportionality (F1 value = 0.197).

The BERT-BiLSTM-CRF model, which was trained on the German Bert model, performed best in predicting proportionality (F1 score = 0.56) and has the best accuracy: on average, the model correctly predicted 90.5% of the labels. However, a look at the recall (= 0.478) shows that even the best model was far from identifying all sentences with a proportionality.

6.1.1 Effect of using BiLSTM and CRF

The experiment showed that the best models for identifying proportionality were all equipped with BiLSTM layers. Among the BERT models, the effect of using LSTM can be observed: Compared to the simple BERT classifier, which can only classify sentences individually, combining BERT with BiLSTM brings a significant improvement. This is in line with our expectations. The additional use of CRF brings only a small change: except for one case,[14] it is a performance improvement. The fact that the improvements by the CRF are so minor can be explained by the fact that we have relatively few documents, and some are very long. The findings align with similar studies (Bhattacharya et al 2023). Overall, the experiment allows us to conclude that using BiLSTM is critical to successful classifications. This is to be expected, as BiLSTMs can look over the entire sequence of sentences during classification. Models that could only look at one sentence during classification performed systematically worse.

6.1.2 Effect of self-trained vs. pre-trained embeddings

However, not all models with BiLSTM perform particularly well: the difference between the BiLSTM models with self-trained embeddings and with pre-trained embeddings is striking. Notably, the models with the newly trained embeddings have the worst F1 scores overall but achieve excellent precision values (especially the BiLSTM-CRF model, which achieves the best precision: 0.963). Thus, there is a trade-off between precision and recall: it identifies very few sentences with proportionality, but these predictions are very reliable. The BiLSTM models with pre-trained sentence vectors show the same trade-off, but in the other direction: they have a high recall (BiLSTM-CRF achieves the best recall: 0.499) but with comparatively low precision. This means the models classify many sentences as having proportionality, but these predictions are less reliable. Nevertheless, the F1 scores of the BiLSTM models with pre-trained sentence embeddings are quite satisfactory and close to those of the BERT models.

The use of pre-trained sentences embedding proved to be helpful. Compared to the tfidf vector and the newly trained embeddings, the pre-trained sentence embeddings performed significantly better in both the SVC and the BiLSTM/BiLSTM-CRF models.

6.1.3 Effect of using BERT models

As mentioned above, the BiLSTM models based on BERT provided the best balance between precision and recall. Relying on BERT models proved to be a successful strategy overall: even among the models trained without LSTM (the baseline models that had to make the prediction based on the sentence alone), the BERT classifier performed

[14] The Accuracy and F1 Scores of the BERT-BiLSTM-CRF are worse than those of the BERT-BiLSTM for those models based on the bert-multilingual.

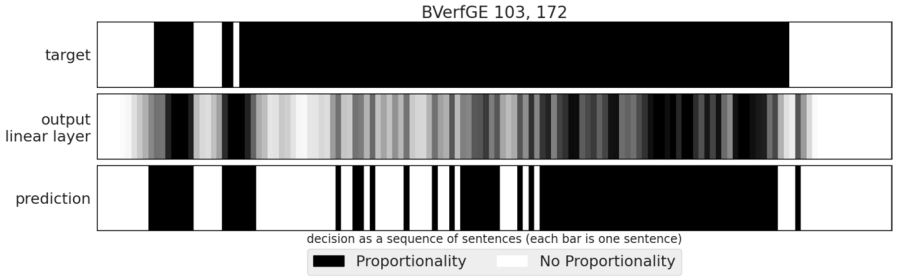

Fig. 3 Graphic visualizing the classification and its errors of the GFCC decision BVerfGE 103, 172

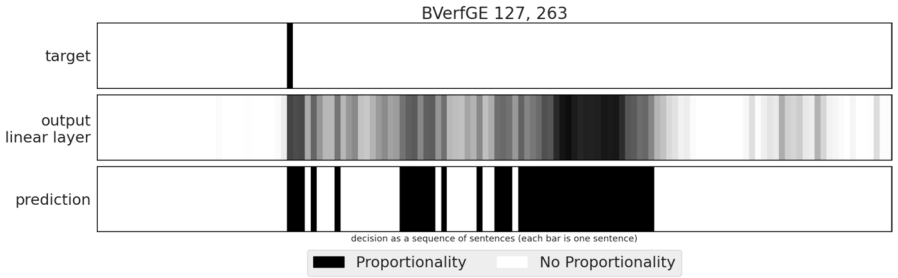

Fig. 4 Graphic visualizing the classification and its errors of the GFCC decision BVerfGE 172, 263

best. Among the LSTM models, they also produced some of the best results, and the BiLSTM-CRF with German BERT was the best-performing model.

The comparison between the two BERT variants shows a clear trend: the German BERT (*dbmdz/bert-base-german-uncased*) produced better results than the multilingual version (*bert-base-multilingual-uncased*) in every configuration. However, a limitation of our study is that we do not have a German legal-BERT. Given the results of other studies, it would be reasonable to expect a further increase in performance (Thalken et al. 2023; Chalkidis et al. 2019). However, using a German BERT is already a significant improvement in classifying German legal sentences.

6.2 Error analysis

In light of the results just discussed, we want to go one step further to understand what makes the classification of proportionality so tricky. As mentioned above, we hypothesized that the challenge for classification is, in particular, to identify all sentences in decisions with long proportionality passages. A hint in this direction is that those models that classify based on individual sentences perform worse than BiLSTM models, which can use the entire decision sequence as information for classification. To better understand the causes of errors, another best-performing

BERT-BiLSTM-CRF model is trained based on *dbmdz/bert-base-german-uncased*. This time, there is only a single stratified sampled 80:20 train-test split.

The test dataset consists of 49 decisions. Of these, 12 decisions contain a proportionality annotation. There are only two misclassified sentences among the 37 decisions without any proportionality annotation. This suggests that most errors occur in decisions with proportionality. To better understand the nature of the problem, two of the decisions with the highest number of misclassifications were selected for further analysis.

Figures 3 and 4 illustrate the decisions. Both figures have an identical structure. They each consist of three elements and illustrate the decision, with a shared X-axis representing the sequence of the texts. The upper part shows the target of the classification, i.e., how the decision is annotated in the data set. Black represents *proportionality*, and white *no proportionality*. The lower part shows the prediction output by the CRF of the model. The middle part shows an intermediate step resulting from the last linear layer before the CRF, showing gradual uncertainties.

Figure 3 shows a decision with a very long proportionality passage. It is easy to see that the model does a very good job of recognizing proportionality at the beginning and end of this passage. However, there are also longer passages where misclassifications are made. A close reading of the decision shows that the legitimate aim of the state measure, the first step of the proportionality test, is disputed in the first part of this passage. This is very plausible for a misclassification. The description of the aim of state measures sometimes is not immediately succeeded by possible other parts of the proportionality test. Other times, however, other parts of the test follow directly after the discussion of the legitimate aim. This is the kind of case we expected. The classification depends on the context and thus becomes difficult for models in long passages.

Figure 4 shows a different problem layout. Only one sentence is annotated in the document.[15] Reading the passages shows that the court states that proportionality could be applied but then refrains from doing so. In an extended passage, the GFCC then discusses the violation of a fundamental right, but proportionality does not play a role. However, the model predicts proportionality resp. considers it very likely. This is not entirely implausible. Human readers would also expect proportionality in a fundamental rights dispute before the GFCC. Nevertheless, the model seems to have a problem of separation.

From this brief analysis of the errors, we can conclude that the challenge lies in the decisions with proportionality — however, the problems there may be quite different. On the one hand, the question of how the model can assign longer passages to proportionality arises. On the other hand, the challenge is that it can clearly distinguish proportionality from simple fundamental rights tests.

[15] The sentence also contains the term 'proportionality'. It translates to: "The general principle of equality imposes varying limits on the legislator depending upon the subject governed and the differentiating elements, ranging from a mere prohibition on arbitrariness to a strict adherence to proportionality requirements.".

7 Conclusion and future research

In this article, we have discussed proportionality in the case law of the GFCC as an object of argumentation mining. We have introduced and conceptualized proportionality as an important argumentation technique. We have argued that such domain-specific argumentation techniques are particularly interesting for argument mining. An important aspect of our work is the extensive dataset we have presented and made available to the scientific community. This dataset contains sentence-level annotations of proportionality in 300 GFCC decisions. Finally, we used this data for the automatic classification. We aimed to classify whether or not a proportionality test was invoked in a sentence. We tested baseline models (including a rule-based approach and a classical SVC) against more complex neural architectures.

The BERT-BiLSTM-CRF model performed best. Overall, the use of BiLSTM layers proved to be very effective. These allow the models to consider the whole sequence of sentences in a decision for classification. Furthermore, the experiment showed that BERT models performed well, as expected, and domain-specific pretrained word embeddings produced decent results. Overall, however, the experiment showed that classifying sentences with proportionality is not easy. Even the best model had a recall below 0.5. A subsequent qualitative analysis of errors showed that there were almost no errors in decisions without proportionality, while there were various sources of errors in decisions with proportionality.

Given these results, we want to conclude by highlighting two promising prospects for further research:

Further and more in-depth use of the data. The data presented in this study are very rich and offer possibilities for many more research efforts. The annotated variables for proportionality are highly sophisticated. Therefore, these categories can be used in further studies, e.g. to analyze the steps of proportionality.

Deepened cooperation between legal research and argument mining. Fundamentally, our work aims to foster the exchange between legal scholarship and argument mining. Both share an interest in arguments. We argue that it is in the interest of argument mining to consider legal argument techniques. In turn, legal scholarship is interested in reliable results from automation. Our study of proportionality is only a first step. The search for further argumentation techniques or even a final typology of legal argumentation may follow.

Appendix A: Original quote of the GFCC from Table 1

Reference: BVerfGE (official collection of GFCC decisions) Volume 68, 272; pages 282–284.

"Im ursprünglichen Gesetzentwurf war die Forderung nach einer bestimmten Qualifikation der Entwurfsverfasser damit begründet worden, dadurch solle im öffentlichen Interesse etwas bewirkt werden, was sich durch die Bauaufsicht nicht erreichen lasse, nämlich eine allgemeine Verbesserung der baulichen Qualität im Hinblick auf Wirtschaftlichkeit, rationelle Gestaltung und Funktionsfähigkeit der

Gebäude, nicht zuletzt aber auch im Hinblick auf die Baukultur (LTDrucks. 8/55 S. 108 f.). Diese Begründung schließt nicht aus, die Einführung einer besonderen Bauvorlageberechtigung darüber hinaus auch mit dem Gesichtspunkt der öffentlichen Sicherheit zu rechtfertigen. Wenn auch die Baugenehmigungsbehörden verpflichtet sind, Bauvorlagen zurückzuweisen, die entgegen den Regeln der Baukunst, aufgrund falscher statischer Berechnungen oder unter Mißachtung baurechtlicher Vorschriften entworfen wurden, so kann der Gesetzgeber doch im Interesse erhöhter Sicherheit und auch zur Entlastung des Baugenehmigungsverfahrens verlangen, daß die erforderlichen Vorlagen bereits von Fachleuten mit entsprechender Vorbildung und Erfahrung angefertigt und verantwortet werden (vgl. dazu BVerfGE 28, 364 [375]; BayVerfGH, BayVBl. 1978, S. 207 [209]; vgl. auch Rasch/Schaetzell, Hessische Bauordnung, in: Die Praxis der Gemeindeverwaltung, F 3 He, S. 293 f.). b) Das vom Gesetzgeber gewählte Mittel, vom Planvorlageberechtigten den Erwerb einer bestimmten fachlichen Qualifikation zu verlangen, war zur Erreichung der gesetzgeberischen Ziele geeignet und erforderlich. Insbesondere ist nicht erkennbar, wie sich diese Ziele durch andere, die Berufsausübung weniger berührende Regelungen hätten erreichen lassen. Die für die Planvorlageberechtigung bei einfacheren Bauvorhaben vorgeschriebene Mindestqualifikation ist ferner nicht unzumutbar hoch, sondern ermöglicht es neben den Architekten und Bauingenieuren auch Handwerksmeistern und den ihnen Gleichgestellten, in erheblichem Umfang im Rahmen ihrer handwerklichen Tätigkeit Bauvorlagen anzuerkennen. Zweifelhaft könnte allenfalls sein, ob sämtliche in § 91 Abs. 4 HBO genannten einfacheren Bauvorhaben nicht nur als genehmigungsbedürftig zu behandeln sind, sondern darüber hinaus von einem planvorlageberechtigten Entwurfsverfasser verantwortet werden müssen. Der Gesetzgeber hat indessen bereits differenziert zwischen genehmigungsbedürftigen Vorhaben, für welche lediglich Architekten und Bauingenieure planvorlageberechtigt sind, anderen genehmigungsbedürftigen einfacheren Bauvorhaben, für welche auch Handwerksmeister und Gleichgestellte Bauvorlagen anerkennen dürfen, ferner nur anzeigebedürftigen (§ 88 HBO) und schließlich genehmigungs- und anzeigefreien Vorhaben (§ 89 HBO sowie die Freistellungsverordnung vom 29. Oktober 1979). Eine weitere Differenzierung kann dem Gesetzgeber von Verfassungs wegen nicht vorgeschrieben werden. Gegen die ihm obliegende nähere Abgrenzung bestehen um so weniger verfassungsrechtliche Bedenken, als die Bauaufsichtsbehörde im Einzelfall bei technisch einfachen baulichen Anlagen auf die Bestellung eines Entwurfsverfassers verzichten kann (§ 77 Abs. 3 HBO)."

Appendix B: creation of sentence embeddings

We created vector representations for sentences based on pre-trained word embeddings (word2vec). The Word Embedding model was created with Gensim, an open-source NLP library (Řehůřek and Sojka 2010). The word embedding model was trained exclusively on the GFCC's case law (Wendel and Möllers 2023). This dataset contains 10665 decisions. The model was trained in 200 epochs, representing 55906 words in 200 dimensions. It is publicly available (Lüders 2024).

The model was used to create a vector representation for each sentence (following Arora et al. 2017): $\vec{P} = \sum_{w \epsilon P} a_w \vec{v}_w$ Each sentence vector $\vec{P}$ is a mean of its word vector $\vec{v}_w$ multiplied by a weight a_w for each word w. The weights $a_w = \frac{\alpha}{\alpha + p_w}$ are indirectly proportional to the frequency of each word in the corpus p_w. The result is a 200-dimensional vector representation for each sentence.

Acknowledgements We thank Christoph Möllers, Ivan Habernal, the editor Kevin Ashley, two anonymous reviewers, and the participants of the CompText Conference 2023 for their helpful comments on previous versions of this paper. We also thank the LLCon project team, our annotators, and Franziska Walz for their support of our project.

Author contributions All authors contributed to the study's conception and design. Material preparation and analysis were performed by Kilian Lüders. Legal expertise was contributed by Bent Stohlmann. The first draft of the manuscript was written by Kilian Lüders and all authors worked on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The work was funded by the DFG Leibniz Prize for Prof Dr Christoph Möllers, LL.M., which was awarded by the public German Research Foundation (Deutsche Forschungsgemeinschaft).

Data availability The Data is available at: Lüders, K., Wendel, L., Reule, S., Stohlmann, B., Hoeft, L., & Tischbirek, A. (2024). Verhältnismäßigkeit—Proportionality. An annotated dataset of GFCC decisions. [Data set]. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.10513684

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article. Note for the purpose of transparency/publication ethics: The article is new and has not been published anywhere else. However, it should be noted that the data presented here will be used for other publications.

References

Arora S, Liang Y, Ma T (2017) A simple but tough-to-beat baseline for sentence embeddings. In: 5th International conference on learning representations. https://openreview.net/forum?id=SyK00v5xx

Artstein R, Poesio M (2008) Inter-Coder Agreement for Computational Linguistics. Comput Linguist 34:555–596. https://doi.org/10.1162/coli.07-034-R2

Bhattacharya P, Paul S, Ghosh K, Ghosh S, Wyner A (2019) Identification of rhetorical roles of sentences in indian legal judgments. arXiv. https://doi.org/10.48550/arXiv.1911.05405

Bhattacharya P, Paul S, Ghosh K, Ghosh S, Wyner A (2023) DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents. Artif Intell Law 31:53–90. https://doi.org/10.1007/s10506-021-09304-5

Brewer S (2018) Interactive virtue and vice in systems of arguments: a logocratic analysis. Artif Intell Law 28:151–179. https://doi.org/10.1007/s10506-019-09257-w

Cabrio E, Villata S (2018) Five years of argument mining: a data-driven analysis. In: Proceedings of the 27th international joint conference on artificial intelligence. International joint conferences on artificial intelligence organization, Stockholm, pp 5427–5433. https://doi.org/10.24963/ijcai.2018/766
Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: the muppets straight out of law school. ArXiv https://doi.org/10.48550/arXiv.2010.02559
Chalkidis I, Kampas D (2019) Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif Intell Law 27:171–198. https://doi.org/10.1007/s10506-018-9238-9
Cohen M, Dahan S, Khern-am-nuai W, Hajime S, Touboul J (2023) The use of AI in legal systems: determining independent contractor vs. employee status. Artif Intell Law. https://doi.org/10.1007/s10506-023-09353-y
Correia F, Alemida A, Nunes JL, Santos K, Hartmann I, Silva F, Lopes H (2022) Fine-grained legal entity annotation: a case study on the brazilian supreme court. Inf Process Manag 59:102794. https://doi.org/10.1016/j.ipm.2021.102794
Costa Y, Oliveira H, Nogueira V, Massa L, Yang X, Barbosa A, Oliveira K, Vieira T (2023) Automating petition classification in brazil's legal system: a two-step deep learning approach. Artif Intell Law. https://doi.org/10.1007/s10506-023-09385-4
Clavié B, Alphonsus M (2021) The unreasonable effectiveness of the baseline: discussing SVMs in legal text classification. arXiv. https://doi.org/10.48550/arXiv.2109.07234
Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://doi.org/10.48550/ARXIV.1810.04805
Fleiss J (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382. https://doi.org/10.1037/h0031619
Grabmair M, Ashley K, Chen R, Sureshkumar P, Wang C, Nyberg E, Walker V (2015) Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. In: Proceedings of the 15th International Conference on Artificial Intelligence and Law, New York, pp 69–78. https://doi.org/10.1145/2746090.2746096
Graves A (2012) Long short-term memory. In: Graves A (ed) Supervised sequence labelling with recurrent neural networks. Berlin: Springer, pp 37–45 https://doi.org/10.1007/978-3-642-24797-2_4
Graves A, Schmidhuber J (2005) framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18:602–610. https://doi.org/10.1016/j.neunet.2005.06.042
Greco C, Tagarelli A (2023) Bringing order into the realm of transformer-based language models for artificial intelligence and law. Artif Intell Law. https://doi.org/10.1007/s10506-023-09374-7
Oliver WM (2020) Transformers for classifying fourth amendment elements and factors tests. In: Serena V, Jakub H, Petr K (eds) Frontiers in artificial intelligence and applications legal knowledge and information systems. IOS Press, Amsterdam
Grimm D (2007) Proportionality in Canadian and German constitutional jurisprudence. UTLJ 57:383–397
Habernal I, Faber D, Recchia N, Bretthauer S, Gurevych I, Döhmann I, Burchard C (2023) Mining legal arguments in court decisions. Artif Intell Law. https://doi.org/10.1007/s10506-023-09361-y
Huscroft G, Miller BW, Webber G (2014) Introduction. In: Huscroft G, Miller BW, Webber G (eds) Proportionality and the Rule of Law. Cambridge University Press, Cambridge, pp 1–18
Lang A (2020) Proportionality analysis by the german federal constitutional court. In: Kremnitzer M, Steiner T, Lang A (eds) Proportionality in action. Cambridge University Press, Cambridge, pp 22–133
Lawrence J, Reed C (2019) Argument mining: a survey. Comput Linguist 45:765–818. https://doi.org/10.1162/coli_a_00364
Leitner E, Rehm G,Moreno-Schneider J (2019) Fine-grained named entity recognition in legal documents. In: Semantic systems. The power of AI and knowledge graphs. Proceedings of the 15th International Conference, 272–87 https://doi.org/10.1007/978-3-030-33220-4_20
Lepsius O (2020) The Standard-Setting Power. In: Jestaedt M, Lepsius O, Möllers C, Schönberger C (eds) The German federal constitutional court. Oxford University Press, Oxford, pp 70–130
Lippi M, Torroni P (2016) Argumentation Mining: State of the Art and Emerging Trends. ACM Trans Internet Technol 16:1. https://doi.org/10.1145/2850417
Lüders, K. BVerfG - Word Embedding, *Zenodo*. https://doi.org/10.5281/zenodo.10908253 (2024)
Lüders K, Wendel L, Reule S, Stohlmann B, Hoeft L, Tischbirek A (2024) Verhältnismäßigkeit - Proportionality. An annotated dataset of GFCC decisions., *Zenodo*. https://doi.org/10.5281/zenodo.10513684

Mochales R, Moens M-F (2009) Argumentation mining: the detection, classification and structure of arguments in text. In: Proceedings of the 12th international conference on artificial intelligence and law, Barcelona, pp 98–107 https://doi.org/10.1145/1568234.1568246

Mochales R, Moens M-F (2011) Argumentation mining. Artif Intell Law 19:1–22. https://doi.org/10.1007/s10506-010-9104-x

Möller K (2015) The global model of constitutional rights. Oxford University Press, Oxford

Möllers C (2013) The three branches: a comparative model of separation of powers. Oxford University Press, Oxford

Möllers C (2020) Legality, legitimacy, and legitimation of the federal constitutional court. In: Jestaedt M, Lepsius O, Möllers C, Schönberger C (eds) The German federal constitutional court. Oxford University Press, Oxford, pp 131–196

Mumcuoğlu E, Öztürk C, Ozaktas H, Koç A (2021) Natural language processing in law: prediction of outcomes in the higher courts of turkey in inf process. Manage 58:102684. https://doi.org/10.1016/j.ipm.2021.102684

Nguyen HT, Phi MK, Ngo XB, Tran V, Nguyen LM, Tu MP (2022) Attentive deep neural networks for legal document retrieval. Artif Intell Law. https://doi.org/10.1007/s10506-022-09341-8

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. arXiv. https://doi.org/10.48550/arXiv.1912.01703

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. JMLR 12:2825–2830

Peters A (2021) A plea for proportionality: a reply to Yun-chien chang and xin dai. ICON 19:1135–1145. https://doi.org/10.1093/icon/moab071

Petersen N (2017) Proportionality and judicial activism: fundamental rights adjudication in Canada. Cambridge University Press, Germany and South Africa

Powers D (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation Int J Mach Learn Tech 2:37–63. https://doi.org/10.48550/ARXIV.2010.16061

Rapp C (2023) Aristotle's Rhetoric, In: Zalta E, Nodelman U (eds) The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2023/entries/aristotle-rhetoric/

Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of LREC 2010 workshop new challenges for NLP frameworks. http://www.fi.muni.cz/usr/sojka/presentations/lrec2010-poster-rehurek-sojka.pdf

Saravanan M, Ravindran B, Raman S (2008) Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In: Proceedings of the 3rd international joint conference on natural language processing. https://aclanthology.org/I08-1063

Šavelka J, Ashley K (2016) Extracting case law sentences for argumentation about the meaning of statutory terms. In: Proceedings of the 3rd workshop on argument mining. Berlin, pp 50–59. https://doi.org/10.18653/v1/W16-2806

Shulayeva O, Siddharthan A, Wyner A (2017) Recognizing cited facts and principles in legal judgements. Artif Intell Law 25:107–126. https://doi.org/10.1007/s10506-017-9197-6

Steiner T, Lang A, Kremnitzer M (2020) Introduction: analyzing proportionality comparatively and empirically. In: Kremnitzer M, Steiner T, Lang A (eds) Proportionality in action. Cambridge University Press, Cambridge, pp 1–21

Steiner T, Netzer L, Sulitzeanu-Kenan R (2022) Necessity or balancing: the protection of rights under different proportionality tests. ICON 20:642–663. https://doi.org/10.1093/icon/moac036

Stohlmann B, Lüders K, Tischbirek A, Wendel L, Hoeft L, Reule S (2024) Konsolidierung statt Siegeszug. Der Staat 63:2

Stone Sweet A, Mathews J (2019) Proportionality balancing and constitutional governance: a comparative and global approach. Oxford University Press, Oxford

Thalken R, Stiglitz E, Mimno D, Wilkens M (2023) Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement. In: Bouamor H, Pino J, Bali K (ed) Proceedings of the 2023 conference on empirical methods in natural language processing. Singapore, pp 9252–9265 https://doi.org/10.18653/v1/2023.emnlp-main.575

Tischbirek A (2020) Die Verhältnismäßigkeitsprüfung: Methodenmigration zwischen öffentlichem Recht und Privatrecht. Mohr Siebeck, Tübingen.

Toulmin S (2003) The uses of argument, Updated. Cambridge University Press, Cambridge

Valvoda J, Cotterell R, Teufel S (2023) On the role of negative precedent in legal outcome prediction. Trans Assoc Comput Linguist 11:34–48. https://doi.org/10.1162/tacl_a_00532

Waltl B, Bonczek G, Scepankova E, Matthes F (2019) Semantic types of legal norms in German laws. Artif Intell Law 27:43–71. https://doi.org/10.1007/s10506-018-9228-y

Weinrib LE (2007) The postwar paradigm and American exceptionalism. In: Choudhry S (ed) The migration of constitutional ideas. Cambridge University Press, Cambridge, pp 84–112

Wendel L, Möllers C (2023) Korpus der Entscheidungen des Bundesverfassungsgerichts (2.0), *Zenodo*, https://doi.org/10.5281/zenodo.10369205

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac O (2020) HuggingFace's transformers: state-of-the-art natural language processing. arXiv. https://doi.org/10.48550/arXiv.1910.03771

Wyner A, Mochales-Palau R, Moens M-F, Milward D (2010) Approaches to text mining arguments from legal cases. In: Francesconi E, Montemagni S, Peters W, Tiscornia D (eds) Semantic processing of legal texts. Springer, Berlin, 60–79 https://doi.org/10.1007/978-3-642-12837-0_4

Zhong L, Zhong Z, Zhao Z, Wang S, Ashley K, Grabmair M (2019) Automatic Summarization of Legal Decisions using Iterative Masking of Predictive Sentences. In: Proceedings of the 17th international conference on artificial intelligence and law. New York, pp 163–72 https://doi.org/10.1145/3322640.3326728

Zhu J, Wu J, Luo X, Liu J (2023) Semantic matching based legal information retrieval system for COVID-19 pandemic. Artif Intell Law. https://doi.org/10.1007/s10506-023-09354-x