Multilabel Legal Element Classification on German Parliamentary Debates in a Low-Ressource Setting

Martin Hock ÷
Technische Universität Dresden
martin.hock@tu-dresden.de

Christopher Klamm ÷
University of Mannheim
klamm@uni-mannheim.de

Abstract

Parliamentary debates provide a broad overview of (legal) pieces of evidence for supporting or opposing the use of force by a state. If a state backs its practice by referring to a legal concept or the legal elements of that concept, the existence of a rule of customary international law (CIL) may be assumed. Traditionally, however, parliamentary debates have rarely been used as a source of CIL. We address this research gap with a joint approach that combines methods from political science, legal studies and natural language processing in order to ascertain the existence of CIL regarding the legal concepts of humanitarian intervention and responsibility to protect. We introduce a new framework and dataset to tackle the task of automatic *legal elements* classification `LegalECGPD` to analyse the use of force in german parliamentary debates. We performed multiple experiments in low-resource settings, showing the need of in-domain expertise and the existing limitations of supervised approaches when faced with tasks necessitating the interpretation of rich contextual information. Our resources are available under an open-source license for further research.

1 Introduction

The use of force by states is unlawful. The Charter of the United Nations (UNC, the most important treaty of international law) prohibits every threat of or use of force (Art. 2 (4) UNC). There are two undisputed exceptions to the prohibition of the use of force: self-defence (art. 51 UNC) and authorization by the Security Council of the United Nations (arts. 39 and 42 UNC). Two further concepts - humanitarian intervention (HI) and responsibility to protect (R2P) - are legally disputed. Arguments supporting the lawfulness of the latter concepts are often based at least partly on customary international law (CIL) (Gray, 2018, p. 40-64). CIL consists of state practice that is accompanied by a sense of legal obligation, the so-called opinio iuris (Lepard, 2010, p. 6-7). State practice and opinio iuris can be found in all branches of the state (International Law Commission, 2018, conclusion 5). All legal concepts are composed of *legal elements*[1], *these are the requirements that have to be fulfilled in order to achieve legal consequences and effects* (Wienbracke, 2013, p. 25-39). This highlights the importance of the legal elements. Legal elements, however, are highly context specific and cannot be assumed by a given word order. They are always composed of the requirements for legal consequences and effects in a given legal rule and vary from rule to rule. In order to prove the existence of opinio iuris the arguments brought forward to substantiate a legal concept are an important factor: if a state backs its practice by referring to a legal concept in general or to the legal elements of that concept the existence of opinio iuris and thus CIL can be assumed (Lepard, 2010, p. 6-7). The legal elements can be found inter alia in parliamentary debates. For example, Ludger Volmer during the `KOSOVO` debate[2]:

> *“ Es kann keinen Zweifel darin geben, daß es überfällig war, den boshaftesten Despoten in Europa [Element 1], der Krieg gegen sein eigenes Staatsvolk führt, es entwurzelt, in die Wälder treibt und ermorden läßt, [Element 2] in seine Schranken zu verweisen, um eine humanitäre Katastrophe noch größeren Ausmaßes zu verhindern. [Element 3]".*

Scholarship on international law largely ignores parliamentary debates as a source of opinio iuris and thus CIL (a notable exceptions is Henckaerts et al. (2005)). It nevertheless refers to national laws that are enacted and the circumstances of their enactment that need to be taken into account when ascertaining opinio iuris (International Law Commission, 2018, conclusion 6). Part of these circumstances are the parliamentary debates that are

÷contribution details in app. F

[1] German: Tatbestandsmerkmale

[2] https://dserver.bundestag.de/btp/13/13248.pdf

Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences, pages 29–45
September 22, 2023. ©2023 Association for Computational Linguistics

conducted in connection with the legislation. In the case of Germany it is highly important to study parliamentary debates since according to the German constitution, parliament is the only body that may legally decide on the use force in international affairs. If parliamentary debates are considered explicitly in international law scholarship the methodological difficulties of including large amounts of text are stressed (see for example Payandeh and Aust (2018, 638) and Bajrami (2022, 160)). We close this gap and treat parliamentary debates as a source of CIL. We provide a new framework for annotating legal elements in parliamentary debates and an annotation of four debates with this new framework, creating our novel "**Legal** **E**lement **C**lassification on **G**erman **P**arliamentary **D**ebates" LegalECGPD dataset. Additional complexity when analysing parliamentary debates is added by the large amount of text in the debates. Legal expertise beyond word search is needed since legal arguments are often ambiguous (i.e. a single sentence can be applied to more than one legal element). Furthermore, the legal concept referred to by a speaker is often not made explicit in parliamentary debates. For example, Minister of Defence Volker Rühe in the KOSOVO debate[3]:

> *"Es geht aber um die Abwehr einer humanitären Katastrophe."*

Implicit in this claim is the legal concept of HI. It is not, however, explicitly mentioned. Nevertheless, the legal element of humanitarian catastrophe is stated. In order to deal with these ambiguities and the lack of explicit references to legal concepts the present system goes beyond word search and shows the need for a more comprehensive approach. From an international law point of view the paper asks weather legal elements can be found in parliamentary debates and thus substantiate the claim that opinio iuris regarding HI and R2P exists. Furthermore, it is asked whether the applied methods of Natural Language Processing (NLP) are sufficiently precise in order to automate the subsumption of parliamentary debates under legal elements.

Advantages in NLP show the possibilities of applying new contextualized language models (Devlin et al. (2019); Reimers and Gurevych (2019); Brown et al. (2020); Lewis et al. (2020); BigScience et al. (2022), and many more) to deal with the automatic identification of supporting and opposing argumentative sentences within natural language (Cabrio and Villata, 2018; Lawrence and Reed, 2019; Reimers et al., 2019; Schaefer and Stede, 2020; Toledo-Ronen et al., 2020; Chakrabarty et al., 2019; Wang et al., 2020; Vecchi et al., 2021; Lapesa et al., 2023). These two fields of research are to be combined in order to enable the analysis of legal elements regarding the validity of legal concepts. This helps international law scholarship to ascertain the opinio iuris of states via the legislature and substantiate the claim to validity of a given legal concept faster and on a broader empirical basis. Analyzing arguments in legal texts, adapting annotation schemes to the legal domain and the overall creation of domain-adapted models is an actively studied NLP area (Haigh, 2018; Yamada et al., 2019; Poudyal et al., 2020; Zhong et al., 2020; Xu et al., 2020; Zhang et al., 2022; Grundler et al., 2022; Chalkidis et al., 2022; Bergam et al., 2022; Niklaus et al., 2023b; Habernal et al., 2023). At the same time, focusing on legal arguments on the intersection between international law and NLP on political texts tends to be rather underexposed in the existing literature. Parliamentary debates can be considered a cross-domain use case inasmuch as they treat questions of international law in an genuinely political setting. As of yet, there are no sufficiently fine-grained analyses regarding legal elements in the context of HI and R2P discussed in debates.

The contributions of our work address several points: First, (1) we introduce a new insight-driven task on the legal element classification in a cross-domain environment. Second, (2) we provide a theoretical-based framework to annotate parliamentary debates and a novel corpus LegalECGPD based on it with 476 sentences (including four debates with 16.836 lines, 324 identified legal elements for 238 sentences), concluded by a legal expert. Furthermore, (3) we present an expert-based analysis of this new corpus giving comprehensive interpretation of the label distribution found. Afterwards, (4) we performed four different state-of-the art deep learning setups in our low-resource setting with transformer-based contextualized sentence embedding and domain-adaptation. Finally, (5) we did a comprehensive error analysis showing multiple limitations of the used models. We conclude that due to the overall moderate performance an expert supported approach is still needed, which points to the need for legal experts in such complex settings.

[3] https://dserver.bundestag.de/btp/13/13248.pdf

2 Related Work

Our work is related to (a) data-driven methods in legal studies as well as international relations scholarship and (b) legal text analysis in NLP.

2.1 Data-driven methods in legal studies

Over the last years the use of data-driven methods in international legal scholarship with several strands of research (Holtermann and Madsen, 2016, p. 11-18; Holtermann and Madsen, 2015) has emerged (Tyler, 2017; Davies, 2020; Dyevre, 2021) leading to the claim of an "empirical turn in international legal scholarship" (Shaffer and Ginsburg, 2012). Empirical legal research aims to identify facts and evidence in order to better understand the topics law regulates and to generate knowledge about the functioning of a given legal system through systematic research supported by quantitative and qualitative data (Eisenberg, 2011, p. 1720; Boom et al., 2018, p. 8; van Dijck et al., 2018). This is where NLP can be used to advance the research agenda. Empirical international legal research has focused on several subfields of international law (Alschner et al., 2017; Ginsburg and Shaffer, 2009), (Posner and de Figueiredo, 2005), (Evangelista and Tannenwald, 2017) and decisions by international courts (Aletras et al., 2016; Medvedeva et al., 2020) or debates in the UN Security Council (Glaser et al., 2022; Patz et al., 2022). In doing so, attention has been given to inter alia big-data analysis or the representation of judicial networks (Coupette, 2019). Research has started to employ machine learning and NLP on subfields of international law (Nay, 2018; Eckhard et al., 2020; Dyevre, 2021; Alschner, 2020)[4]. Nevertheless, there are gaps in the research. Even though questions of customary international law have been singled out as being able to benefit from empirical and digital methods, not much research has been conducted (Megiddo, 2019). Questions regarding the prohibition on the use of force as well as parliamentary debates have been mostly excluded in international law scholarship (a notable exception is Lewis et al. (2019)) Analysis of parliamentary debates and the use of force using empirical methods are conducted in political science and international relations scholarship (Vignoli (2020); Wagner (2020); Hock (2021)) but largely excluded from legal scholarship. Thus, the present work is set on the interdisciplinary boundaries between international legal scholarship, international relations scholarship and NLP.

2.2 Legal Elements Classification in Natural Language Processing

We introduced that *legal elements* prove the existence of legal concepts. Therefore, we can conceptualize these elements close to the concept of *arguments supporting or opposing (as premises or reasons)* (Lawrence and Reed, 2019) the existence of an *implicitly or explicitly claimed legal concept (claim).*

Automated argument mining, a rapidly emerging subfield of Natural Language Processing (NLP), finds wide application in the automatic detection, verification, and characterisation of arguments (Lippi and Torroni, 2016; Cabrio and Villata, 2018; Stede and Schneider, 2018; Lawrence and Reed, 2019; Vecchi et al., 2021; Lapesa et al., 2023). The benefits of new contextualized models for argument mining have been exhibited in the recent research (Reimers et al., 2019; Wang et al., 2020; Habernal et al., 2023).

More and more studies are now focusing on legal texts. This trend is driven by two factors: creating automated systems to process legal text can reduce the repetitive and time-consuming tasks of legal practitioners and scholars. Moreover, these systems can offer a reliable reference to those not familiar with the legal domain (Zhong et al., 2020).

The existing body of research has made available legal corpora that serve as the subject of various classification tasks, thereby giving rise to new datasets (Zhang et al., 2022). Chalkidis et al. (2023) have offered a multinational English legal corpus consisting of 11 sub-corpora that encompass legislation and case law from six English-speaking legal systems, namely, the EU, the Council of Europe, Canada, the US, the UK, and India. A recent release by Niklaus et al. (2023b) includes a multilingual legal corpus spanning 24 languages (including German, English, Spanish, and others) from 17 jurisdictions.

Specific to argumentation mining, Poudyal et al. (2020) have made available an annotated corpus composed of decisions from the European Court of Human Rights (ECHR), building on the previously annotated corpus provided by Mochales and Moens (2011). Grundler et al. (2022) released a corpus

[4] For a recent data set on argument mining and the European Court of Human Rights see Poudyal et al. (2020); see also Altwicker (2019) and Barczentewicz (2021) for an overview of the methodological challenges for international law scholarship.

for argument mining, composed of decisions of the Court of Justice of the European Union. Other examples include Japanese judgement documents (Yamada et al., 2019) and case holdings on legal decisions (Zheng et al., 2021). Habernal et al. (2023) recently introduced a labeled corpus for argument mining based on the English corpus of the European Court of Human Rights.

Legal language is often categorized as a "sublanguage". Like other specialized domains such as medical texts, legal texts (laws, pleadings, contract) possess distinctive properties such as specialized vocabulary, formal syntax, and semantics rooted in extensive domain-specific knowledge. This leads to unique properties in comparison to generic corpora (Haigh, 2018). A base model like BERT (Devlin et al., 2019) often falls short in specialized domains (Beltagy et al., 2019). In this regard, Chalkidis et al. (2020) suggested LegalBERT, pre-trained on multiple legal corpora such as EURLEX and LEGISLATION.GOV.UK. Another study by Zheng et al. (2021) employed the complete English Harvard Law case corpus to pretrain CaseLaw-BERT. Legal language models have been also pre-trained for Italian (Licari and Comandè, 2022), Romanian (Masala et al., 2021), and Spanish (Gutiérrez-Fandiño et al., 2021) as well. Furthermore, Niklaus et al. (2023b) trained a multilingual Legal-XLM and evaluated it on the newly introduced LEXTREME (Niklaus et al., 2023a), a legal benchmark. Habernal et al. (2023) performed continuous pre-training on the ECHR corpus using the RoBERTa-Large model for argument mining.

Moreover, argument mining in political debates, particularly in (German) parliamentary debates, remains rather unexplored. Limited research has been conducted in this area, including works by Menini et al. (2018), who annotated speeches by Nixon and Kennedy during the 1960 Presidential campaign, Visser et al. (2021), who annotated the 2016 US presidential debates, and Hüning et al. (2022), who used messages from an online survey about a Local Rent Control Initiative for argument mining. Another noteworthy contribution is by Mestre et al. (2021), who built a corpus consisting of labeled sentence pairs from the 2020 US political election debates. Recently, Mancini et al. (2022) released a multi-modal corpus, where the text input is enriched by and aligned to the audio input.

Unlike existing research, our legal context is situated in the cross-domain of parliamentary debates *(political texts)*, limiting the usability of existing methods due to the differing use of language. This demonstrates that a clear separation between legal texts and other types of texts does not always represent reality. Therefore, the cross-domain task of legal element recognition on German parliamentary debates is not covered by the existing datasets and models, which means that we can not use existing corpora or models to adapt them to our use case. Instead, we need to provide a *new corpus* that represents legal element types in the context of *German parliamentary debates* more comprehensively.

3 Legal Background

In this paper we address several unique types of legal elements derived from the legal concepts of HI and R2P. Both share the same argumentative basis: gravest violations of human rights may serve as a justification for the use of force by third states. They lack, however, a clear-cut distinction from each other as well as clear dogmatic legal grounding. Both have often been based on CIL as well as expansive interpretations of the UNC. By the end of the 19th century HI was mostly considered lawful if there was a just cause for intervention. In principle, if a state conducted gross abuses against its population, any other state that was willing to intervene militarily in order to stop these abuses had the right to do so (Neff, 2005, p. 217-218). Thus, the two legal elements of a humanitarian catastrophe and the protection of locals were needed. With the signing of the UNC, the idea of HI became superseded by art. 2 (4). While there have been some uses of force under the justification of HI between 1945 and the early 1990s, the claim to the legality of HI remained weak (Dave, 2009, p. 37-38; Gray, 2018, p. 40-44). The failure to prevent the genocide in Rwanda in 1994 and NATO's intervention without an authorization by the Security Council under the framework of HI in Kosovo in 1999 lead to renewed debate regarding the lawfulness of HI (Crossley, 2018, p. 418-420; Thakur, 2016). These discussions culminated in the development of R2P. R2P was brought forward by the International Commission on Intervention and State Sovereignty and argued for two major changes. Contrary to HI, R2P's main focus is not the right to intervene but the protection of the population. Additionally, sovereignty is seen as conditional to the protection of a population from suffering serious harm. If a state is

Label	Legal Element	Definition	Example
HUMA	**Humanitarian catastrophe**	The code is used if the speaker refers to a humanitarian catastrophe taking place or being imminent (major human rights violations that amount to war crimes, genocide, ethnic cleansing and crimes against humanity) that makes the use of force necessary.	The situation in this country is a humanitarian catastrophe, people starve and suffer, therefore we must use force to stop the aggressor.
PROT	**Protection of local civilians**	This code applies if the speaker considers that the need to protect the local civilians from major human rights violations, that amount to war crimes, genocide, ethnic cleansing and crimes against humanity makes the use of force necessary.	We have to use our country's military to protect the civilians in this country.
FAIL	**Failure to Protect by home state**	This code applies if the speaker considers that the home state has failed to protect its population from war crimes, genocide, ethnic cleansing and crimes against humanity makes the use of force necessary.	This country is not protecting its people from the crimes against humanity occurring, thus we need to use our military.
LAST	**Last Resort**	This code applies if the speaker considers to the use of force as a last resort and that all peaceful means (such as diplomacy) are exhausted.	We have tried every diplomatic means available but to no avail, there is no choice but to use force.
PROP	**Proportionality of the use of force to the threat**	This code applies if a speaker sees the way force is used in a proportional manner (including that civilians are protected as far as possible or receive special treatment to help with the suffering.)	When we use force we take every possible precaution to protect the civilians from our attacks.
REAS	**Reasonable prospect of success**	This code applies if the speaker argues that a reasonable prospect of success is given.	Using the military is always risky but we are sure that we will succeed.
AUTH	**Rightful authority given**	This code applies if the speaker argues that a rightful or legitimate authority for the use of force is given (this includes but is not limited to references to the Security Council).	We have every right to use force and our actions are covered by the Security Council.
INTE	**Right intention**	This code applies if the speaker refers to having the right intention of the use of force. This might be the case, for example, if speakers refer to a moral cause for going to war as being given.	This is not a war for our national interest it is a moral duty.

Table 1: Framework adapted from codebook from Hock (2021), drawing on work from Wagner (2020).

not willing or not able to protect its population, the responsibility for doing so shifts to the international community (Bellamy, 2014, p. 1-3; Saba and Akbarzadeh, 2018, p. 244-245). In order for the responsibility to pass on to the international community, several criteria have to be fulfilled: a just cause has to be given, the use of force has to be conducted with the right intention as a last resort, proportional to the threat, and with reasonable chance of success. The authority to authorize an intervention under R2P should generally lie with the Security Council (ICISS, 2001, p. XI-XIII). The 2005 World Summit Outcome[5] endorsed the R2P in principle. It also stressed the sovereignty of states and the importance of an authorization by the Security Council. This brought R2P closer in line with traditional understandings of the UNC. R2P could be invoked in cases of war crimes, genocide, ethnic cleansing, and crimes against humanity but only if a state manifestly failed to protect its population. Nevertheless, a definite and exhaustive list of criteria needed for a situation of R2P was not brought forward. Following from the above, several legal elements that are shared between HI and R2P can be distilled: humanitarian catastrophe, protection of locals, failure to protect by the home state, right intention, last resort, proportionality of the use of force to the threat, reasonable chance of success, rightful authority given. R2P remains controversial amongst states and the international community (Crossley, 2018). Thus, the analysis of customary international law and state's opinions regarding HI and R2P remains highly relevant.

4 Dataset for Legal Elements Classification

We claimed that a new dataset is needed to cover the task of classifying legal elements in parliamentary debates. In this section, we will give details on the creation of our "LegalECPD-dataset". (1) Inspired by Hock (2021), drawing on work from Wagner (2020), we introduce an adapted framework to annotate the different facets of legal elements regarding HI and R2P. (2) We then show the annotation process of four debates (KOSOVO (BTP 13/248), LIBYA (BTP 17/095), SYRIA-A (BTP 18/042), SYRIA-B (BTP 18/044). Moreover, (3) we analyze the generated dataset LegalECGPD of the identified 324 legal elements in 238 sentences in terms of the distribution and characteristics of the debates.

Dataset Creation. We base our work on four parliamentary debates regarding the authorization to

[5] UN Doc. A/Res/60/1 (24. October 2005) para. 138-140

			multilabel								
set	*ratio*	*sentences #*	Huma	Prot	Fail	Last	Prop	Reas	Auth	Inte	labels #
full	100%	**238**	51	52	20	44	8	32	39	78	*324*
			stratified-debate split								
train	*70%*	**165**	31	38	13	30	5	23	27	52	*219*
dev	*10%*	**25**	6	8	3	3	1	3	7	8	*39*
test	*20%*	**48**	14	6	4	11	2	6	5	18	*66*
			cross-debate split								
train	KOSOVO, SYRIA-A	**191**	46	43	13	36	7	29	24	64	*262*
dev	LIBYA	**20**	4	3	4	4	0	1	10	3	*29*
test	SYRIA-B	**28**	1	6	3	4	1	2	5	11	*33*

Table 2: Summary of our legal elements dataset, detailing the *sample counts* and *multilabel distributions* for two data splitting strategies: the **stratified-debate split** (ensuring balanced representation of the four debates in each set) and the **cross-debate split** (allocating two debates for training, and one each for development and testing).

.

use force[6]. Parliamentary debates are especially interesting for two reasons. Firstly, they are a means to ascertain the opinio iuris of a state. Secondly, while they cover questions of international law, they do so as part of a political speech, not as for example a legal analysis. Thus, legal concepts will be mentioned but intertwined with genuinely political arguments. It is safe to assume that the legal elements mentioned may be vague or ambiguous due to the political nature of the texts. In order to analyse the legal basis of these concepts we take the cases of the war in Kosovo, the war in Libya and the war against ISIS (Syria-A and Syria-B) as examples. The datasets draws on all German debates authorizing to use of force for the first time (the ISIS debates are strongly connected in the sense that Syria-A is the debate that continues to the vote in Syria-B. The latter is thus considered as to fall into these criteria as well). This is based on the observation that the war in Kosovo was a catalyst for the creation of R2P. The war against Libya as well as the war against ISIS can be considered as examples in which R2P featured prominently - even though in both cases other legal justifications, such as authorization by the Security Council and self-defence played a more important role. *No further debates* that fulfil the criteria of authorizing the use of force for the first time in a situation in which either HI or R2P might apply exist. For example, while there are more parliamentary debates on the use of force in Kosovo, the situation has changed from an international law perspective. After the Kosovo War the debate turned towards the presence of foreign forces in the Kosovo and the fulfillment of inter alia Security Council resolution 1244[7]. Thus, an expansion of the dataset on the ground of the above mentioned criteria is not possible.

Annotation. Our coding framework is adapted from the coding scheme of Hock (2021), which draws on the work of Wagner (2020). We adapt the coding scheme to better cover the concepts of HI and R2P. This includes merging and expanding several codes that aimed for grasping different facets of legal theory into codes adapted to cover legal elements. For example "just cause", "just war", and "right intention, warfare as morally justified" have been merged to "right intention". The codes "failure to protect by home state" and "proportionality of the use of force to the threat" have been included in our coding framework. Furthermore, codes that focused exclusively on questions of legal theory, such as for example "just war" or "state of exception makes legality less important" have been deleted. Our aim is to include all legal elements made for the legal framework of HI and R2P. We used the presented annotation scheme and defined categories (”codebook”). Our codebook itself consists of seven domain-specific categories. We did our study with one legal domain expert. Our legal

[6] https://dserver.bundestag.de/btp/{ID}.pdf (ID=13/13248|17/17095|18/18042|18/18044)

[7] UN Doc. S/Res/1244 (1999)

expert is a 30-35 years old (male) with a strong academic background in international law and more than three years experience in this domain. He checked more than $15k$ sentences of all debates.

Statistics and Analysis. The resulting dataset (Tab. 2) contains 324 legal elements in 238 sentences (only 1-2% of all sentences in the debates included legal elements). While there are many similarities (App. D), such as the large amount of right intention and protection of local civilians in all debates - as was expected since we are analysing the same legal exception to the prohibition of the use of force - there are several noteworthy aspects. The element of humanitarian catastrophe is most widely used in the debate on `Kosovo`. This illustrates the point that HI was re-discovered as a legal doctrine with the war in `Kosovo`. The shift towards the R2P explains the relative decline in the usage of humanitarian catastrophe. This is connected with the element of right intention being used most frequently in the `Kosovo` debate as well. Since the use of force in support of human rights was a rather novel occurrence in the `Kosovo` war, this finding is not surprising. `Syria-A` features protection of local civilians prominently, `Syria-B` right intention. Regardless of several outliers, the distribution pattern is comparable over cases. From an international law perspective, it is regrettable that only around 1.5 per cent of all lines contained legal elements. Normatively, this questions whether a strong base for opinio iuris can be found in parliamentary debates at all. Further research is necessary to determine the relative strength of the arguments made. Thus, the lack of relevant lines could be due to a strong belief that the legal basis is clear and unequivocal. Speakers may not have referred towards legal elements because they took their existence for granted.

5 Automatic Approach for Legal Elements Classification

In this section, we analyse the performance of NLP methods on legal element classification on German parliamentary debates. We focuses on the concrete type of the legal element, performing a multi-label classification task in a few-shot setting.

Task. We model the *classification of legal elements* as a sentence-level multi-label classification task. Given a sentence s composed of words w_i, where $i \in \{1, ..., n\}$, the goal is to assign a list of legal elements $e = (e_0, ..., e_m)$ to the sentence. For example[8]:

> *"Es kann keinen Zweifel darin geben, daß es überfällig war, den boshaftesten Despoten in Europa [right intention], der Krieg gegen sein eigenes Staatsvolk führt, es entwurzelt, in die Wälder treibt und ermorden läßt, [failure to protect] in seine Schranken zu verweisen, um eine humanitäre Katastrophe noch größeren Ausmaßes zu verhindern. [humanitarian catastrophe]"* $\xrightarrow{e}$ **`Inte, Fail, Huma`**

Dataset. We use our novel `LegalECGPD` dataset and create two different dataset splits (Tab. 2): *stratified random split* and *cross-debate split*. In the *stratified random split*, we randomly but stratified the dataset, allocating 70% for train, 10% for development, and 20% for test. This ensured a balanced representation of our different debates across the subsets. In the *cross-debate split*, designed for cross-debate evaluation, we divided the data based on the perspectives captured in the debates. Two debates were included in the train set, one in the development set, and the remaining one in the test set. This approach facilitated the exploration of distinct perspectives across different debates, enabling a comprehensive evaluation of the model's performance in a cross-debate scenario.

Models. We base our analyses on freely available traditional and state-of-the-art NLP methods. We intend to demonstrate the performance of existing models to provide a basis for further research. We applied four types of models (A) dictionary-base (**DB**), (B) feature-based (**FB**), (C) transformer-based fine-tuning (**FT**) and (D) domain-adapted sentence-transformer (**DASent**):

(A) Dictionary-based (DB): We apply *dictionary-based models* with two pre-defined lexicons: an expert-curated one with domain knowledge (App. B.1) and a generated lexicon via Pointwise Mutual Information (PMI) (Church and Hanks, 1989) for statistical associations. These models offer a simple and efficient baseline.

(B) Feature-based (FB): Furthermore, we test *feature-based models*, extracting specific features from data for classifications. We used TF-IDF to measure word importance (Sparck Jones, 1988) and GermanBERT embeddings (Chan et al., 2020) to represent data and capture semantic relationships. We use a multi-layer perceptron (Rumelhart et al., 1986) as classification head incorporating the features for classification (App. B.2)

[8] https://dserver.bundestag.de/btp/13/13248.pdf

(C) Transformer-based Fine-tuning (FT): We also test a transformer-based GermanBERT (Chan et al., 2020) with fine-tuning on our task-specific dataset (App. B.3). This approach leverages prior knowledge from the BERT base model.

(D) Domain-Sent-Transformer (DASent): Finally, we apply SetFit (Tunstall et al., 2022) that adapts sentence-transformer models to our domain by training on `LegalECGPD` (App. B.4). SetFit employs contrastive learning for fine-tuning. This technique distinguishes between similar and dissimilar sentence pairs to capture semantic relationships. The adapted model generates domain-specific sentence embeddings for classification and is "efficient [...] for few-shot fine tuning." (ib.)

Results. Our results (Tab. 3 and App. 6) show that domain adaptation using sentence embeddings outperforms other approaches (**DASent**, .63 F1-Micro, ±.01 std). Specifically, the task-specific fine-tuning method **(FT)** shows comparable performance (.61 F1-Micro, ±.01 std) to the domain-adaptation technique, albeit slightly lower. On the other hand, the baseline models, including the dict-based **(DB)** and feature-based models **(FB)**, demonstrate inferior performance. These findings emphasize the effectiveness of leveraging domain-specific knowledge encoded within sentence embeddings for improved performance in the given task.

Model	**F1micro**	*Pre*	*Rec*	**F1macro**	*Pre*	*Rec*
DB-Expert	.16	*.75*	*.06*	.09	*.19*	*.06*
DB-PMI	.17	*.24*	*.14*	.16	*.19*	*.18*
FB-TFIDF	.43	*.58*	*.35*	.31	*.45*	*.25*
FB-BERT	.55	*.64*	*.55*	.45	*.58*	*.43*
FT-BERT	.61	***.72***	*.53*	.48	*.60*	*.44*
DASent (SetFit)	**.63**	*.57*	***.70***	**.63**	***.65***	***.71***

Table 3: Results on our *stratified random test set*.

Model	**F1micro**	*Pre*	*Rec*	**F1macro**	*Pre*	*Rec*
DB-Expert	.00	*.00*	*.00*	.00	*.00*	*.00*
DB-PMI	.19	*.25*	*.16*	.14	*.18*	*.14*
FB-TFIDF	.37	*.50*	*.29*	.22	*.33*	*.19*
FB-BERT	.38	*.38*	*.38*	.36	*.40*	*.38*
FT-BERT	.47	***.64***	*.37*	.34	*.44*	*.33*
DASent (SetFit)	**.57**	*.47*	***.71***	**.43**	***.43***	***.56***

Table 4: Results on our *cross-debate test set* `SYRIA-B`.

Additionally, when considering the more challenging dataset that involved cross-debate evaluation, our results (Tab. 4 and app. 6) continue to showcase the effectiveness of domain adaptation using sentence embeddings (**DASent**, .57 F1-Micro ±.01 std).

Error Analysis. Our results of the best performing model `DASent` show that the most interesting task is connected to the label `INTE`, covering the legal element of right intention. Here, we have seen several divergences between the coder and the model. This is due to the fact that `INTE` covers arguments that are deeply intertwined with moral judgments. Furthermore, in these cases, connections between different parts of the argument are often implicit and highly dependent on context as well as prior knowledge. They represent fringe cases that are difficult to evaluate even with domain expertise (translations DE →EN in app. E).

“Nennen Sie mir einen weltweit, der sich mehr darum bemüht, dass dieser politische Prozess zustande kommt." (id 173)

Here, the model labeled the argument as `LAST`, the domain expert as `Last` and `INTE`. While it is clear that the argument centers partly around the use of force as being an action of last resort, the argument is also morally based. This is due to the fact that the speaker claims to be the one who is the most concerned with keeping the peace. Another example is:

“Es kann keinen Zweifel darin geben, daß es überfällig war, den boshaftesten Despoten in Europa, der Krieg gegen sein eigenes Staatsvolk führt, es entwurzelt, in die Wälder treibt und ermorden läßt, in seine Schranken zu verweisen, um eine humanitäre Katastrophe noch größeren Ausmaßes zu verhindern.“ (id 138)

Here, the moral judgment is made via the classification of the political leader as evil despot instead of simply as enemy. Implicit in the understanding of the leader as evil despot (as well as explicit in the further parts of the text) is the humanitarian catastrophe and the failure to protect. Thus, the domain expert has labeled this as `HUMA`, `FAIL`, and `INTE` while the model did label it as `HUMA` and `PROT`. Arguably, the label `PROT` could be used as well, nevertheless, `FAIL` is the more fitting label. Moral judgments are also contained in the following argument:

“Wenn wir diese schrecklichen Szenen als Fernsehzuschauer in Westeuropa einfach konsumieren würden, ohne zu handeln, dann würden wir letztlich mit einer rostigen Rasierklinge unser Gesicht zerschneiden und unser eigenes Gesicht entstellen.” (id 72)

Here, the reference towards a rusty razor cutting one’s own face can be understood as a moral judgement claim. It was thus labeled as `INTE` by the domain expert. The model labeled it as `PROT`. Here, however, the argument refers towards the self-image and their own understanding of moral and ethical considerations and not towards the protection of civilians as such. Another example is:

> *“Wir können nicht tatenlos zusehen, wenn sich regionale Faustrechte entwickeln und Menschenrechte in Regionen so verletzt werden, daß es zu humanitären Katastrophen kommen kann, weil das Gewaltmonopol der Vereinten Nationen nicht ausgeübt werden kann.“ (id 114)*

In this last example the model labeled it as HUMA and PROT. The domain expert labeled it as HUMA and INTE. Here, the notion of taking the law into one's own hand leads to the argument centering around legality and the moral role of the law. Nevertheless, there is merit to the label PROT. It remains to be said that there are far-reaching consequences if the existence of a rule of CIL that considers HI and R2P to be lawful were ascertained. Ultimately, warfare might occur more often (Orford, 2003).

6 Conclusion

International legal concepts are in most cases based on treaty law or CIL. CIL consists of state practice and opinio iuris. We claim that opinio iuris can be found in parliamentary debates. A legal concept consists of legal elements. Thus, in order to prove the existence of opinio iuris one has to find legal elements in parliamentary debates. We tried to ascertain the existence of opinio iuris regarding HI and R2P by analysing legal elements in parliamentary debates. Our use case offers a cross-domain approach inasmuch as it combines two domains that usually are treated separately, i.e. legal elements in genuinely political texts. This presents a novel task for NLP methods. We offer a new dataset and a contribution to the fields of NLP as well as empirical legal scholarship. Our experiments have shown several results. There is a surprisingly low amount of legal elements mentioned in parliamentary debates. The distribution of legal elements follows expected patters inasmuch as all cases cover the same legal concepts. Nevertheless, it became evident that the used NLP models do not provide sufficient accuracy (yet) in such a few-shot multilabel setting. Thus, domain specific knowledge is needed. Our provided framework enables future research and our data set is available under an open-source license[9].

7 Future Work

Possible future work on this project could benefit from several different extensions. From the viewpoint of NLP and its methods three major expansions would be beneficial: First, improving the granularity of annotation to extend it to a span- and token-level would potentially yield greater detail and precision in data labeling. Furthermore, the reliability and diversity of annotations could be bolstered by expanding the number of legal experts involved in annotating legal elements. This approach may augment the range and depth of perspectives, thereby enhancing the overall quality and balance of the dataset. Second, building on the recently introduced models for multilingual legal language, there is an opportunity to develop a cross-domain language model specifically tailored for analyzing legal language in political debates. Lastly, conducting a more in-depth study to interpret the domain features that the model uses to classify legal elements would be beneficial. This could involve a legal expert-guided analysis of the typical underlying features that should be deployed, and further expanding the model to concentrate more on these conceptual features. The ultimate objective is to develop a model that is guided more by defined legal concepts than merely by linguistic characteristics, potentially leading to more accurate and relevant interpretation and classification of legal elements. From the viewpoint of international law two further extensions would be of value to further work. In order to further ascertain the CIL-base of HI and R2P it would be beneficial to study other parliamentary democracies besides Germany since the more states support a legal concepts based on CIL the more substantial is the claim towards the legality of that concept. Furthermore, the present methods could be applied to legal concepts besides HI and R2P in the area of international law and the use of force, such as for example expansive understandings of the right to self-defence. This would provide international law scholarship and international political decision making with an empirically grounded substantiation of the legality of the use of force within the context of legally contested situations.

8 Limitations

The present paper shows several problems when dealing with international law empirically. We showed the challenging aspects of classification in a low resource setting. Unfortunately due to specific use case we can not simply scale up the amount of data. As discussed our dataset covers the use case for the legal domain. Furthermore, le-

[9] github.com/chkla/LegalECGPD

gal norms have a inherent ambiguity regarding the situation they are applicable to. More often than not, a legal argument is multi-faceted and highly complex with often implicit and highly context-based references as well as moral judgments. Only rarely will a speaker invoke clearly which legal norm he might be referring to. Furthermore, the legal concepts of HI and R2P remain difficult to define. Consequently, a statement made in a debate may count towards more than one legal element. Thus only a limited amount of the argumentative depth of a legal argument can be covered with the present methods. This points towards a general problem of empirical legal sciences that needs to be answered in further research.

9 Ethics Statement

When developing a model to predict legal element types in parliamentary debates, several ethical considerations must be addressed. *Firstly*, the accuracy and reliability of the model are critical; moderate results might indicate that the model may not capture the nuances and complexities of legal concepts within debates. Involving legal experts in the training and evaluation process can help to refine the model, ensuring that it properly reflects legal terminologies and concepts. However, expert involvement should be balanced to avoid biases that may inadvertently be introduced by the experts. Incorporating the fact that only one expert was involved adds another layer of ethical consideration. Relying on a single expert could introduce a lack of diversity in perspectives and potentially skew the model towards the biases and opinions of that particular individual. Legal interpretation often requires a range of perspectives to account for the complexities and subtleties of language and context. With just one expert, there is a risk that the model may not be as robust or as representative as it could be with the input from multiple experts with diverse backgrounds and areas of expertise. Nevertheless, even the inclusion of more than one legal expert might not lead to significantly better results. Legal questions are always based on interpretation. This interpretation cannot be fully objective, even though the methodology of legal science aims to reduce the subjectivity involved in interpretation. Thus, standard solutions such as taking the average of several experts might improve performance but do not in each and every case lead to better results. Additionally, domain expertise is scarce. It is therefore difficult to include more than one expert *Secondly*, the transparency and explainability of the model are essential, particularly in the legal domain where the decisions and analyses can have far-reaching consequences. Adding to this aspect concerning the transparency and explainability of the model, it is crucial to consider that error analysis typically reveals only a fraction of the possible errors. This limitation in understanding the full scope of the model's errors is a vital ethical concern. It means that the model could have underlying issues that are not immediately apparent, and these unidentified issues could lead to incorrect or misleading predictions. *Thirdly*, one must consider the potential misuse of the model. If the model is not highly accurate, relying on its predictions without human verification could lead to misinterpretations of legal elements in debates, which in turn could have policy implications or affect legal interpretations and decisions.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93.

Wolfgang Alschner. 2020. Sense and Similarity: Automating Legal Text Comparison. In Ryan Whalen, editor, *Computational Legal Studies: The Promise and Challenge of Data-Driven Legal Research*, pages 9–28. Edward Elgar.

Wolfgang Alschner, Joost Pauwelyn, and Sergio Puig. 2017. The Data-Driven Future of International Economic Law. *Journal of International Economic Law*, 20(2):217–231.

Tilmann Altwicker. 2019. International Legal Scholarship and the Challenge of Digitalization. *Chinese Journal of International Law*, 18(2):217–246.

Shpetim Bajrami. 2022. *Selbstverteidigung gegen nichtstaatliche Akteure*. Mohr Siebeck.

Mikolaj Barczentewicz. 2021. Teaching Technology to (Future) Lawyers. *Erasmus Law Review*, 15(1).

Alex J. Bellamy. 2014. *The Responsibility to Protect*. Oxford University Press.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Noah Bergam, Emily Allaway, and Kathleen Mckeown. 2022. Legal and political stance detection of SCOTUS language. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 265–275, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale

Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Willem Boom, Pieter Desmet, and Peter Mascini. 2018. Empirical legal research: charting the terrain. In Willem Boom, Pieter Desmet, and Peter Mascini, editors, *Empirical Legal Research in Action. Reflections on Methods and their Applications*, pages 1–22.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5427–5433. ijcai.org.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Corinna Coupette. 2019. *Juristische Netzwerkforschung*. Mohr Siebeck GmbH and Co. KG.

Noele Crossley. 2018. Is R2P still controversial? Continuity and change in the debate on 'humanitarian intervention'. *Cambridge Review of International Affairs*, 31(5):415–436.

Benjamin Dave. 2009. LAST RESORT: BRIDGING PROTECTION AND PREVENTION. *International Journal on World Peace*, 26(4):37–62.

Gareth Davies. 2020. The Relationship between Empirical Legal Studies and Doctrinal Legal Research. *Erasmus Law Review*, 13(2):3–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gijs van Dijck, Shahar Sverdlov, and Gabriela Buck. 2018. Empirical Legal Research in Europe: Prevalence, Obstacles, and Interventions. *Erasmus Law Review*, 11(2):105–119.

Arthur Dyevre. 2021. Text-mining for Lawyers: How Machine Learning Techniques Can Advance our Understanding of Legal Discourse. *Erasmus Law Review*, 15(1):7–23.

Steffen Eckhard, Ronny Patz, and Mirco Schönfeld. 2020. Keine Spur von Sprachlosigkeit im Sicherheitsrat. *Vereinte Nationen*, 68(5):219.

Theodore Eisenberg. 2011. The Origins, Nature, and Promise of Empirical Legal Studies and a Response to Concerns. *University of Illinois Law Review*, (5):1713–1738.

Matthew Evangelista and Nina Tannenwald, editors. 2017. *Do the Geneva Conventions Matter?* Oxford University Press.

Tom Ginsburg and Gregory C. Shaffer. 2009. How Does International Law Work: What Empirical Research Shows. In Peter Cane and Herbert Kritzer, editors, *The Oxford Handbook of Empirical Legal Research*, pages 753–748. Oxford University Press.

Luis Glaser, Ronny Patz, and Manfred Stede. 2022. UNSC-NE: A Named Entity Extension to the UN Security Council Debates Corpus. *Journal for Language Technology and Computational Linguistics*, 35(2):51–67.

Christine Gray. 2018. *International Law and the Use of Force*. Oxford University Press.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in CJEU decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish legalese language model and corpora. *ArXiv preprint*, abs/2110.12201.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*.

Rupert Haigh. 2018. *Legal English*. Routledge, Abingdon, Oxon; New York, NY.

Jean-Marie Henckaerts, Louise Doswald-Beck, Carolin Alvermann, Knut Dörmann, and Baptiste Rolle. 2005. *Customary International Humanitarian Law*. Cambridge University Press.

Martin Hock. 2021. The Influence of Strategic Culture on Legal Justifications. *Erasmus Law Review*, 14(2):68–81.

Jakob V. H. Holtermann and Mikael Madsen. 2015. European New Legal Realism and International Law: How to Make International Law Intelligible. *Leiden Journal of International Law*, 28(2):211–230.

Jakob V. H. Holtermann and Mikael Madsen. 2016. What is Empirical in Empirical Studies of Law? A European New Legal Realist Conception. *iCourts Working Paper Series*, 77.

Hendrik Hüning, Lydia Mechtenberg, and Stephanie Wang. 2022. Detecting arguments and their positions in experimental communication data. *SSRN Electronic Journal*.

ICISS. 2001. *The Responsibility to Protect. Report of the International Commission on Intervention and State Sovereignty*. International Development Research Centre, Ottawa, ON, Canada.

International Law Commission. 2018. *Draft conclusions on identification of customary international law, with commentaries*. United Nations.

Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. 2023. Mining, assessing, and improving arguments in NLP and the social sciences. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dubrovnik, Croatia. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Brian D. Lepard. 2010. *Customary International Law*. Cambridge University Press, Cambridge.

Dustin A. Lewis, Naz K. Modirzadeh, and Gabriella Blum. 2019. Quantum of Silence: Inaction and Jus ad Bellum. In *Harvard Law School Program on International Law and Armed Conflict*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Daniele Licari and Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*, volume 3256 of *CEUR Workshop Proceedings*, Bozen-Bolzano, Italy. CEUR. ISSN: 1613-0073.

Marco Lippi and Paolo Torroni. 2016. Argumentation Mining. *ACM Transactions on Internet Technology*, 16(2):1–25.

Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torroni. 2022. Multimodal argument mining: A case study in political debates. In *Proceedings of the 9th Workshop on Argument Mining*, pages 158–170, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, 28(2):237–266.

Tamar Megiddo. 2019. Knowledge Production, Big Data and Data-Driven Customary International Law. *SSRN Electronic Journal*.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22.

John Nay. 2018. Natural Language Processing and Machine Learning for Law and Policy Texts. *SSRN Electronic Journal*.

Stephen C. Neff. 2005. *War and the Law of Nations*. Cambridge University Press.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023a. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *ArXiv preprint*, abs/2301.13126.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023b. Multilegalpile: A 689gb multilingual legal corpus. *ArXiv preprint*, abs/2306.02069.

Anne Orford. 2003. *Reading Humanitarian Intervention. Human Rights and the Use of Force in International Law*. Cambridge University Press.

Ronny Patz, Manfred Stede, and Luis Glaser. 2022. Die Wahl der Worte im Sicherheitsrat. *Vereinte Nationen*, 70(6):260.

Mehrdad Payandeh and Helmut Philipp Aust. 2018. Praxis und Protest im Völkerrecht. *JuristenZeitung*, 73(13):633.

Eric A. Posner and Miguel F. P. de Figueiredo. 2005. Is the International Court of Justice Biased? *The Journal of Legal Studies*, 34(2):599–630.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Arif Saba and Shahram Akbarzadeh. 2018. The Responsibility to Protect and the Use of Force: An Assessment of the Just Cause and Last Resort Criteria in the Case of Libya. *International Peacekeeping*, 25(2):242–265.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Gregory Shaffer and Tom Ginsburg. 2012. The Empirical Turn in International Legal Scholarship. *American Journal of International Law*, 106(1):1–46.

Karen Sparck Jones. 1988. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. In *Document Retrieval Systems*, pages 132–142. Taylor Graham Publishing, GBR.

Manfred Stede and Jodi Schneider. 2018. Argumentation Mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Ramesh Thakur. 2016. Rwanda, Kosovo, and the International Commission on Intervention and State Sovereignty. In Alex J. Bellamy and Tim Dunne, editors, *The Oxford Handbook on the Responsibility to Protect*, pages 94–113. Oxford University Press.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. *ArXiv preprint*, abs/2209.11055.

Tom R. Tyler. 2017. Methodology in Legal Research. *Utrecht Law Review*, 13(3):130.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

Valerio Vignoli. 2020. Where are the doves? Explaining party support for military operations abroad in Italy. *West European Politics*, 43(7):1455–1479.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. Annotating argument schemes. *Argumentation*, 35:101–139.

Wolfgang Wagner. 2020. *The Democratic Politics of Military Interventions*. Oxford University Press.

Hao Wang, Zhen Huang, Yong Dou, and Yu Hong. 2020. Argumentation mining on essays at multi scales. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5480–5493, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mike Wienbracke. 2013. *Juristische Methodenlehre*. C.F. Müller, Heidelberg ; München ; Landsberg ; Frechen ; Hamburg.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Huihui Xu, Jaromír Šavelka, and Kevin D. Ashley. 2020. Using argument mining for legal text summarization. In *Frontiers in Artificial Intelligence and Applications*, volume 334: Legal Knowledge and Information Systems.

Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. *Artificial Intelligence and Law*, 27(2):141–170.

Gechuan Zhang, Paul Nulty, and David Lillis. 2022. A decade of legal argumentation mining: Datasets and approaches. In *Natural Language Processing and Information Systems*, pages 240–252, Cham. Springer International Publishing.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Appendix

B Models

In this section, we provide further details regarding the models employed in our experimental setting:

B.1 Expert-Based Dictionary

We created an *expert-based dictionary*, drawing upon the wealth of knowledge from our domain expert. In the following table, we present the various terms and expressions contained within our dictionary (Tab. 5).

B.2 Feature-based Classification with MLP heads

We take the sentence embedding from the pre-trained models to perform a classification using a small *Multi-Layer Perceptron* (MLP) with 32 hidden layers (alpha=$1e$-5, random_state=$\{42, 111, 133\}$). We place these methods as classification headers over the TFIDF and GermanBERT embeddings (Chan et al., 2020)[10] to perform legal element classification.

[10] https://huggingface.co/deepset/gbert-base

labels	words
`Huma`	"humanitäre", "katastrophe", "gewalt", "massaker", "notlage"
`Pro`	"gewalt", "massaker", "notlage", "vertreibung", "flüchtlinge", "mord", "kriegsverbrechen", "opfer"
`Fail`	"angriffe", "bevölkerung", "regierung", "präsident", "bürgerkrieg", "vertreibung", "kriegsverbrechen", "volk", "säuberungen", "staatsterror"
`Last`	"letztes", "äußerstes", "mittel", "ultima", "ratio", "lösung", "politisch", "allein", "gewalt", "militärisch"
`Prop`	"begrenzt", "vertretbar", "proportional", "luftschlag", "erforderlich", "phasen", "angemessen", "gleichwertig"
`Reas`	"erfolg", "aussicht", "vertretbar", "lösung", "glaubhaft", "wirkung", "realistisch", "chance", "erreichbar", "ziel"
`Auth`	"sicherheitsrat", "resolution", "autorisierung", "staatengemeinschaft", "un", "vereinte nationen", "generalsekretär", "nato", "mandat", "eu"
`Inte`	"moral", "freiheit", "demokratie", "menschenrechte", "friede", "friedlich", "diplomatisch", "lösung", "tyrann", "stabilität"

Table 5: Dictionary.

B.3 Transformer-based Fine-Tuning

We use pre-trained models trained on monolingual GermanBERT (Chan et al., 2020). We fine-tuned the model with the HuggingFace *transformers* library (Wolf et al., 2020) on the random split (epochs=20, lr=1e-5, epsilon=2e-08 and batch=8) and the debates split (epochs=15, lr=5e-5, epsilon=1e-08 and batch=8) with three different seeds $\{42, 111, 133\}$ and selected the best performing model based on the evaluation loss.

B.4 Domain-Sent-Embeddings (DASent)

We used the framework SetFit in our experiments (Tunstall et al., 2022). We trained the domain-adapted sentence embeddings with SetFit for 10 epochs and 20 iterations (on three different seeds with batch_size=16, learning_rate=2e-5, warmup_proportion=0.1). SetFit employs a method known as contrastive learning in the fine-tuning process of the sentence transformer (Reimers and Gurevych, 2019). Contrastive learning is a natural language processing method in which the model learns to distinguish between a pair of sentences that are similar (positive pair) and a pair that are dissimilar (negative pair). The model learns to bring positive pairs closer in the embedding space while simultaneously pushing negative pairs further apart. This results in a better differentiation between positive and negative pairs. Contrastive learning is capable of capturing meaningful semantic relationships between texts, which significantly improves the model's performance. This is especially beneficial when dealing with small datasets. After this step, the fine-tuned model generates sentence embeddings that are used to train a classification head (Tunstall et al., 2022).

C Negative Example

Find below an example that does not contain any legal elements. This example serves to illustrate the kind of content that is not relevant for legal element classification. A negative example for classification is the following statement made by Karsten D. Voigt in the Kosovo debate[11]:

> *"Es gebührt auch dem bisherigen Kanzler und dem künftigen Kanzler, den bisherigen und künftigen Ministern, die durch diese Art des Zusammenwirkens einen Beitrag zur politischen Kultur in Deutschland geleistet haben, nachdrücklich Dank."* → **Element 1, Element 2, Element3**

No legally relevant statement has been given in this sentence.

D Dataset Characteristics

The following figure provides an overview of the label distribution across all debates, offering insights into the overall composition of the dataset (Fig. 1).

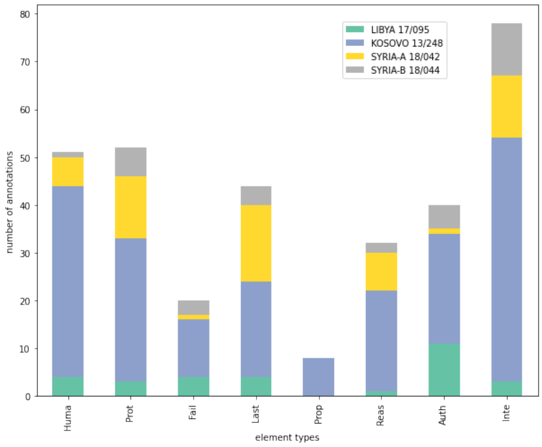

Figure 1: Dataset characteristics, showing the *overall label distribution* (see Tab. 1) for all debates.

E Translation

The translation (DE→EN) has been conducted by the authors:

(**DE**, *id 138*) "Es kann keinen Zweifel darin geben, daß es überfällig war, den boshaftesten Despoten in Europa, der Krieg gegen sein eigenes

[11] https://dserver.bundestag.de/btp/13/13248.pdf

				multilabel																							
				HUMA			PROT			FAIL			LAST			PROP			REAS			AUTH			INTE		
Model	Set	HL	CE	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Stratified Random Split																											
DB-Expert	dev	.20 ± .00	8.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00
DB-Expert	test	.16 ± .00	7.42 ± .00	1.0 ± .00	.36 ± .00	.53 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.50 ± .00	.09 ± .00	.15 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00
DB-PMI	dev	.23 ± .00	6.92 ± .00	.50 ± .00	.33 ± .00	.40 ± .00	.50 ± .00	.12 ± .00	.20 ± .00	.33 ± .00	.67 ± .00	.44 ± .00	.17 ± .00	.33 ± .00	.22 ± .00	.17 ± .00	1.0 ± .00	.29 ± .00	.50 ± .00	.33 ± .00	.40 ± .00	.75 ± .00	.43 ± .00	.55 ± .00	1.0 ± .00	.12 ± .00	.22 ± .00
DB-PMI	test	.22 ± .00	7.33 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.60 ± .00	.27 ± .00	.37 ± .00	.17 ± .00	.50 ± .00	.25 ± .00	.25 ± .00	.33 ± .00	.29 ± .00	.20 ± .00	.20 ± .00	.20 ± .00	.29 ± .00	.11 ± .00	.16 ± .00
FB-TFIDF	dev	.17 ± .00	6.44 ± .20	.50 ± .00	.17 ± .00	.25 ± .00	.69 ± .04	.29 ± .06	.41 ± .06	1.0 ± .00	.33 ± .00	.50 ± .00	.61 ± .08	.67 ± .00	.63 ± .04	.00 ± .00	.00 ± .00	.00 ± .00	.83 ± .24	.33 ± .00	.47 ± .05	1.0 ± .00	.38 ± .07	.55 ± .07	.48 ± .11	.25 ± .1	.33 ± .11
FB-TFIDF	test	.16 ± .00	5.91 ± .18	.87 ± .03	.60 ± .03	.70 ± .01	.10 ± .02	.17 ± .00	.13 ± .01	.00 ± .00	.00 ± .00	.00 ± .00	.62 ± .03	.30 ± .04	.41 ± .05	.00 ± .00	.00 ± .00	.00 ± .00	.58 ± .31	.17 ± .00	.25 ± .04	.43 ± .05	.40 ± .00	.41 ± .02	1.0 ± .00	.41 ± .05	.58 ± .05
FB-GermanBERT	dev	.21 ± .01	5.92 ± .11	.50 ± .00	.17 ± .00	.25 ± .00	.63 ± .04	.50 ± .00	.54 ± .02	1.0 ± .00	.33 ± .00	.50 ± .00	.67 ± .00	.67 ± .00	.67 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	1.0 ± .00	.33 ± .00	.50 ± .00	1.0 ± .00	.29 ± .00	.44 ± .00	.50 ± .00	.25 ± .00	.33 ± .00
FB-GermanBERT	test	.14 ± .00	5.10 ± .21	.81 ± .07	.57 ± .00	.67 ± .02	.24 ± .03	.56 ± .08	.34 ± .04	.44 ± .08	.42 ± .12	.43 ± .10	.64 ± .10	.21 ± .04	.32 ± .06	.00 ± .00	.00 ± .00	.00 ± .00	.59 ± .07	.39 ± .08	.46 ± .06	.70 ± .07	.87 ± .09	.77 ± .03	.91 ± .07	.69 ± .07	.78 ± .03
FT-GermanBERT	dev	.13 ±.01	5.64 ±.14	1.0 ±.00	.22 ±.08	.36 ±.10	.71 ±.06	.42 ±.12	.52 ±.11	.67 ±.47	.22 ±.16	.33 ±.24	.83 ±.24	.55 ±.16	.67 ±.19	.00 ±.00	.00 ±.00	.00 ±.00	.89 ±.16	.56 ±.16	.66 ±.12	1.0 ±.00	.52 ±.07	.68 ±.06	.78 ±.02	.46 ±.06	.58 ±.05
FT-GermanBERT	test	.12 ±.00	5.12 ±.07	.92 ±.01	.81 ±.07	.86 ±.04	.33 ±.09	.44 ±.16	.37 ±.11	.83 ±.24	.42 ±.12	.52 ±.11	.66 ±.00	.36 ±.15	.45 ±.13	.00 ±.00	.00 ±.00	.00 ±.00	.55 ±.42	.17 ±.14	.24 ±.18	.57 ±.00	.80 ±.00	.67 ±.00	.93 ±.06	.56 ±.14	.70 ±.09
DASent (SetFit)	dev	.13 ± .01	4.45 ± .10	.72 ± .21	.33 ± .00	.45 ± .04	.74 ± .02	.71 ± .06	.72 ± .04	.56 ± .08	.67 ± .00	.60 ± .04	.35 ± .07	.67 ± .00	.45 ± .06	.00 ± .00	.00 ± .00	.00 ± .00	.60 ± .00	1.0 ± .00	.75 ± .00	.91 ± .06	.95 ± .07	.93 ± .06	.64 ± .05	.71 ± .06	.67 ± .00
DASent (SetFit)	test	.14 ± .003	4.22 ± .11	.84 ± .03	.86 ± .00	.85 ± .01	.21 ± .01	.67 ± .00	.32 ± .01	.62 ± .03	.67 ± .12	.63 ± .04	.76 ± .01	.58 ± .04	.65 ± .03	.00 ± .00	.00 ± .00	.00 ± .00	.55 ± .03	.67 ± .00	.60 ± .02	.44 ± .03	.93 ± .09	.59 ± .04	.78 ± .03	.67 ± .05	.72 ± .04
Cross-Debate Split																											
DB-Expert	LIBYA	.19 ± .00	8.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00
DB-Expert	SYRIA-B	.16 ± .00	8.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00
DB-PMI	LIBYA	.19 ± .00	7.65 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.50 ± .00	.20 ± .00	.29 ± .00	.00 ± .00	.00 ± .00	.00 ± .00
DB-PMI	SYRIA-B	.19 ± .00	6.81 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.33 ± .00	.50 ± .00	.40 ± .00	.67 ± .00	.40 ± .00	.50 ± .00	.40 ± .00	.18 ± .00	.25 ± .00
FB-TFIDF	LIBYA	.23 ± .01	7.30 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.39 ± .08	.25 ± .00	.30 ± .02	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.14 ± .00	.33 ± .00	.20 ± .00
FB-TFIDF	SYRIA-B	.15 ± .00	6.32 ± .20	.00 ± .00	.00 ± .00	.00 ± .00	.75 ± .00	.50 ± .00	.60 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.47 ± .05	.50 ± .00	.48 ± .03	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	1.0 ± .00	.20 ± .00	.33 ± .00	.39 ± .09	.30 ± .11	.34 ± .10
FB-GermanBERT	LIBYA	.21 ± .01	5.98 ± .37	.44 ± .08	.25 ± .00	.32 ± .02	.00 ± .00	.00 ± .00	.00 ± .00	1.0 ± .00	.25 ± .00	.40 ± .00	.39 ± .28	.25 ± .20	.30 ± .23	.00 ± .00	.00 ± .00	.00 ± .00	.25 ± .00	1.0 ± .00	.40 ± .00	.69 ± .04	.23 ± .05	.35 ± .06	.20 ± .00	.33 ± .00	.25 ± .00
FB-GermanBERT	SYRIA-B	.17 ± .01	5.59 ± .44	.44 ± .08	1.0 ± .00	.61 ± .08	.35 ± .05	.50 ± .00	.41 ± .04	.00 ± .00	.00 ± .00	.00 ± .00	.33 ± .00	.25 ± .00	.29 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	1.0 ± .00	.83 ± .24	.89 ± .16	.67 ± .00	.40 ± .00	.50 ± .00	.44 ± .03	.45 ± .07	.45 ± .05
FT-GermanBERT	LIBYA	.16 ±.02	6.62 ±.27	.44 ±.31	.33 ±.24	.38 ±.27	.10 ±.11	.11 ±.16	.10 ±.13	.00 ±.00	.00 ±.00	.00 ±.00	.44 ±.42	.17 ±.12	.23 ±.17	.00 ±.00	.00 ±.00	.00 ±.00	.00 ±.00	.00 ±.00	.00 ±.00	.72 ±.04	.27 ±.05	.39 ±.06	1.0 ±.00	.67 ±.00	.80 ±.00
FT-GermanBERT	SYRIA-B	.12 ±.01	5.94 ±.33	.61 ±.28	1.0 ±.00	.72 ±.21	.76 ±.17	.61 ±.08	.67 ±.10	.00 ±.00	.00 ±.00	.00 ±.00	.67 ±.47	.17 ±.12	.27 ±.19	.00 ±.00	.00 ±.00	.00 ±.00	.00 ±.00	.00 ±.00	.00 ±.00	1.0 ±.00	.40 ±.00	.57 ±.00	.52 ±.08	.45 ±.07	.48 ±.06
DASent (SetFit)	LIBYA	.14 ± .01	4.12 ± .13	.78 ± .16	.75 ± .00	.76 ± .08	.30 ± .02	.67 ± .00	.41 ± .02	1.0 ± .00	.33 ± .12	.49 ± .13	.36 ± .02	.67 ± .12	.47 ± .05	.00 ± .00	.00 ± .00	.00 ± .00	.67 ± .24	1.0 ± .00	.78 ± .16	.93 ± .05	.87 ± .05	.90 ± .04	.48 ± .03	1.0 ± .00	.64 ± .03
DASent (SetFit)	SYRIA-B	.16 ± .01	3.99 ± .11	.22 ± .02	1.0 ± .00	.36 ± .03	.53 ± .02	1.0 ± .00	.69 ± .02	.00 ± .00	.00 ± .00	.00 ± .00	.33 ± .00	.50 ± .00	.40 ± .00	.00 ± .00	.00 ± .00	.00 ± .00	.83 ± .24	.50 ± .00	.61 ± .08	1.0 ± .00	.60 ± .00	.75 ± .00	.51 ± .03	.88 ± .04	.64 ± .04

Table 6: Performance comparison of our applied models for legal element classification on German parliamentary debates using our two different dataset splits: *Stratified Random Split* and *Cross-Debate Split* (**H**amming **L**oss, **C**overage **E**rror, **Pre**cision, **Rec**all, **F1**-Macro and ± **std**).

Staatsvolk führt, es entwurzelt, in die Wälder treibt und ermorden läßt, in seine Schranken zu verweisen, um eine humanitäre Katastrophe noch größeren Ausmaßes zu verhindern.".

(**EN**, *id 138*) *There can be no doubt, that it was due to put checks on the most evil despot in Europe who wages war against his own people, displaces them, and drives them into the woods in order to avoid a humanitarian catastrophe of even larger extent.*

(**DE**, *id 173*) "Nennen Sie mir einen weltweit, der sich mehr darum bemüht, dass dieser politische Prozess zustande kommt"

(**EN**, *id 173*) *Show me one person in the world, who is more concerned that this political process will take place.*

(**DE**, *id 72*) "Wenn wir diese schrecklichen Szenen als Fernsehzuschauer in Westeuropa einfach konsumieren würden, ohne zu handeln, dann würden wir letztlich mit einer rostigen Rasierklinge unser Gesicht zerschneiden und unser eigenes Gesicht entstellen."

(**EN**, *id 72*) *If we were to simply consume these horrible scenes on the TV screen in Western Europe without acting it would equal to cutting our own face with a rusty razor and disfiguring our own face.*

(**DE**, *id 114*) "Wir können nicht tatenlos zusehen, wenn sich regionale Faustrechte entwickeln und Menschenrechte in Regionen so verletzt werden, daß es zu humanitären Katastrophen kommen kann, weil das Gewaltmonopol der Vereinten Nationen nicht ausgeübt werden kann."

(**EN**, *id 114*) *We cannot stand-by idly when regionally the law is taken into their own hands and in certain regions human rights are violated that much that a humanitarian catastrophe takes place because the monopoly of violence of the United Nations cannot be enforced.*

F Contributions

This collaborative project combines expertise from the fields of political science, law, and natural language processing. The first author produced the underlying taxonomy and theoretical foundation of the project, as well as facilitating the annotation of legal elements in political debates (Section 3). A key part of the project was a collaboration between the authors (Sections 1-2 and 6-9). Furthermore, the coding was done by the first author (Section 4). The second author prepared the dataset for annotation and led the annotation process (Section 4). In addition, the second author designed and implemented various natural language processing tools to automatically predict legal elements in political debates (Section 5). In the subsequent phase, the first author did an error analysis of the predictions generated by the model (Section 5). The remaining aspects of the project were performed in collaboration with each other.