Structured Persuasive Writing Support in Legal Education: A Model and Tool for German Legal Case Solutions

Florian Weber
University of Kassel / GER
weber@uni-kassel.de

Thiemo Wambsganss
EPFL / CH
thiemo.wambsganss@epfl.ch

Seyed Parsa Neshaei
EPFL / CH
seyed.neshaei@epfl.ch

Matthias Söllner
University of Kassel / GER
soellner@uni-kassel.de

Abstract

We present an annotation approach for capturing structured components and arguments in legal case solutions of German students. Based on the appraisal style, which dictates the structured way of persuasive writing in German law, we propose an annotation scheme with annotation guidelines that identify structured writing in legal case solutions. We conducted an annotation study with two annotators and annotated legal case solutions to capture the structures of a persuasive legal text. Based on our dataset, we trained three transformer-based models to show that the annotated components can be successfully predicted, e.g. to provide users with writing assistance for legal texts. We evaluated a writing support system in which our models were integrated in an online experiment with law students and found positive learning success and users' perceptions. Finally, we present our freely available corpus of 413 law student case studies to support the development of intelligent writing support systems.

1 Introduction

Writing persuasive texts plays a major role in law education (Kosse and Butle Ritchie, 2003). As a part of their training for learning how to write legal opinions, law students are typically challenged to solve legal problems or case studies in the form of persuasive case solutions (Enqvist-Jensen et al., 2017). To write a persuasive legal case solution, students in German law courses must be able to follow the structural requirements of the appraisal style (see Figure 1) (Stuckenberg, 2020) and justify their derived conclusions argumentatively via legal claims and premises (see Section 2). To learn a skill such as writing a persuasive case solution, individual feedback is important to the learning process (Hattie and Timperley, 2007; Black and Wiliam, 2009). Individualized feedback for students during their writing or learning processes is lacking, particularly in the field of law. The characteristic large-scale learning scenarios in legal studies, which result in a low supervision ratio, are part of the reason for this. Organizational restrictions are another cause of the absence of personal feedback. For instance, there aren't enough lecturers who can assess students' case solutions (Henderson, 2003). At the same time, technical solutions that could help students improve their legal writing fall short of expectations (Beurskens, 2016). One promising solution to better support students in their writing process and to overcome the limitations in law courses would be the use of writing support systems that could provide individualized feedback to students (Wambsganss et al., 2020a).

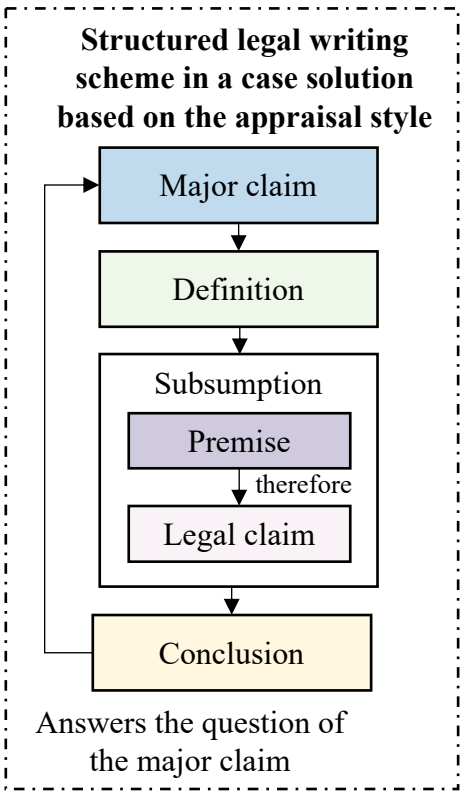

Figure 1: Annotation scheme for structured legal writing in a case solution based on the appraisal style: *major claim, definition, subsumption* and *conclusion*.

Findings of the Association for Computational Linguistics: ACL 2023, pages 2296–2313
July 9-14, 2023

To model legal persuasive writing with predictive algorithms, high-quality annotated corpora are needed. Pioneering work in argumentation mining has already focused on jurisprudence (Mochales and Moens, 2008; Mochales and Ieven, 2009), since the structural approach of legal writing facilitates the unambiguous determination of argumentation components (Lytos et al., 2019; Urchs et al., 2020). Existing corpora in law range from classification of judgments (Urchs et al., 2020), to summarization of legal texts (Hachey and Grover, 2005) and to evaluation of jury verdicts (Poudyal et al., 2019). Corpora dealing with the annotation of structural elements in student written legal texts are not available. A few corpora are suitable for designing and developing systems to support persuasive writing (Stab and Gurevych, 2017b; Wambsganss et al., 2020b; Lawrence and Reed, 2019; Stab and Gurevych, 2014). However, these corpora are of limited use for modeling the structure of writing and argumentation in law, since persuasive writing in the legal domain follows a particular logic (see Section 2) that is not represented by available corpora. Consequently, there is a lack of evaluated annotation schemes and linguistic corpora for training models that support users in legal writing.

Therefore, we propose a novel annotation scheme for persuasive student-written case solutions. We introduce a corpus of 413 student-written case solutions with 25,103 sentences that are annotated for the components of the appraisal style, arguments (legal claim and premises), the relations of the arguments, and the relations of distinct components of the appraisal style. We trained different types of models (e.g. BERT and DistilBERT) and compared their accuracy to analyze which model performs best. Finally, we embedded the three best performing transformer-based BERT models in a novel writing support system that provides individual feedback and recommendations in a writing scenario. The design of our writing support system is based on the theory of learning from errors (Metcalfe, 2017) and aims to provide students with individual feedback on their errors during the writing process (Fazio and Marsh, 2009). We tested the systems in an online learning scenario with law students. The students were asked to use the system to write a case solution. We show promising results in terms of the students' understanding of the appraisal style and their perception of the system. The participants perceive the system as useful and rate the system's feedback as accurate. Our analyzed results support that our presented corpus and the models are able to support students' learning effectively.

Our work makes five major contributions. First, we derive a novel modeling approach for a new data domain by developing an annotation scheme based on the theory of structured legal writing based on the appraisal style (Man, 2022; Stuckenberg, 2020). Second, we present an annotation study based on 100 student case solutions to show that annotation of student case solutions is accurately possible. Based on the annotation, we trained three transformer-based BERT models (Devlin et al., 2019) to demonstrate that the prediction of the annotated structures is possible with a certain accuracy. Fourth, we provide a corpus of 413 student case solutions in German collected in different law lectures. Finally, we show in an online experiment that the models can be used effectively in a writing support system. Therefore, we encourage further investigation into the enhancement of law students' persuasive structured writing and the development of writing support systems using NLP. This research aims to enhance students' skills regardless of their location, time constraints, or instructor availability.

2 Related Work

Persuasive Writing in Law Courses Classically, students are asked to solve legal problems or case studies in the form of persuasive case solutions (Enqvist-Jensen et al., 2017). In these case solutions, students are forced to use specialized and highly concept-driven knowledge. The theoretical knowledge specializes more in the correct application of paragraphs and the setting of priorities in the case solution. In contrast, the concept-driven knowledge is largely composed of the concepts of writing case solutions in a structured way. To do this, students must follow established legal concepts. Among the most important concepts in German jurisprudence are the appraisal style and the judgment style, whereby the appraisal style is primarily important for legal education (Stuckenberg, 2020; Urchs et al., 2020). Since the term "*appraisal style*" is a peculiarity of the German legal language, there is no direct equivalent in English. We define the term appraisal style as "*the form and writing style of a legal opinion*" (Stuckenberg, 2020). The appraisal style is used to solve complex legal

problems. The four elements of appraisal style are briefly explained in Table 1 and supplemented by an example in Figure 2.

Corpora in the Legal Field Although law is a promising discipline for annotating the components of legal writing and arguments due to its fixed logical structure (Moens et al., 2007; Urchs et al., 2020), evaluated open-access corpora for law are rare (Reed, 2006; Mochales and Moens, 2011; Urchs et al., 2020). There are, however, some publicly accessible corpora. Hachey and Grover (2005) present a corpus of 188 annotated English court opinions. To construct a system for automatic summarizing of court judgments, they annotated rhetorical status, significance, and linguistic markup. Other annotated corpora deal explicitly with the annotation of argumentation structures in court decisions (Houy et al., 2013) or legal cases (Mochales-Palau and Moens, 2007). Mochales-Palau and Moens (2007) present a corpus of English-language judicial cases gathered from the European Court of Human Rights (ECHR). They chose 55 papers at random, which included 25 court decisions and 29 admissibility reports. The texts were annotated and studied systematically in two layers (argumentative and non-argumentative sentences). A following study showed that the detection of argumentative sentences in court decisions is possible. Work such as that of Walker et al. (2014) has focused on identifying successful and failed patterns of reasoning in U.S. Court decisions. Patterns of reasoning are identified and used to illustrate the difficulty of developing a type or annotation system for characterizing these patterns. The corpus is based on legal cases of vaccine-injury compensations. There are several German corpora in addition to the largely English-language corpora for recognizing decisions and legal cases. Urchs et al. (2020) created a corpus based on Bavarian Court of Justice decisions. They discover argumentation structures in judgments using 200 court decisions. Other research groups focused on the identification of arguments in the German Federal Constitutional Court's Decision Corpus (Houy et al., 2013) and the development of a German referent corpus comprised of articles from legal journals, decision texts, and norm texts (Gauer et al., 2016).

A number of corpora have previously been proposed in research to enhance students' structured and persuasive writing in real-world applications, including Stab and Gurevych (2017a) and Stab and Gurevych (2014). Stab and Gurevych (2014) produced a corpus based on student essays for building and implementing systems to promote persuasive writing for adaptive feedback using argumentation mining (AM) approaches. Further research uses the corpus as a model to annotate persuasive writings (Carlile et al., 2018) or construct a model for assessing persuasive essays (Ke et al., 2018). However, the existing literature does not adequately transfer corpora for structured writing or reasoning to other educational domains, like law or to other languages.

To summarize, we see that literature falls short of annotated corpora, which can be used to model components in student-written legal case solutions. Without the availability of these corpora, the design of adaptive NLP-based applications for lawful writing is naturally hindered. To the best of our knowledge, there is only one approach by Urchs et al. (2020) that aims to detect the components of legal writing, but the approach focuses on court decisions and the judgment style. Therefore, we aim to address this literature gap by presenting and evaluating an annotation scheme as well as an annotated corpus built on student-written texts with the objective of developing an intelligent writing support system for students in law courses.

3 Construction of the Corpus

3.1 Data Source

The data for our corpus were collected in a law courses at a German university. We compiled the corpus with the case solutions of law students who have written solutions to different legal problems (four different case studies) from different areas of law. In total, we collected 413 legal case solutions, with a typical length of 55.07 sentences and 331.35 tokens per document.[1] The case studies are mainly based on example cases from civil law and are oriented towards basic cases of Musielak and Hau (2005). Students solved the cases as a component of a comprehensive law lecture, utilizing them as a means of exam preparation. It is important to note that the quality of the 413 student-written case solutions may vary, as the students are not all at the same level of proficiency or understanding. The data were collected in the mentioned lecture between 2020 and 2022. The course deals with the teaching of the basics of legal writing and the funda-

[1]The data collection was conducted according to the ethical guidelines of our university.

mental knowledge of business law were introduced. Accordingly, the course has dealt with essential basics that are also important for non-law students, such as business students. The data collected are thus relevant not only in the context of foundational legal education but also for many other German-language legal studies programs (e.g., law courses in the education of business students).

3.2 Annotation Scheme

The correct application of structured legal writing in the appraisal style is the basis for a persuasive legal opinion. In the following, the components of the legal writing structure, as well as its annotation, are explained. The structure consists of four components: *major claim, definition, subsumption* (premise and legal claim), and *conclusion* (Sieckmann, 2020; Backer, 2009) (see Table 1).

Components	Definition
Major claim	The major claim explains the elements of the offense (fact) that are to be fulfilled. It raises a question or possible consequence. The question is discussed in the following steps and is finally answered in the conclusion.
Definition	The definition determines the constituent elements that must occur in the legal problem so that the case solution can come to a conclusion. The elements always depend on the question raised in the major claim.
Subsumption (premise and legal claim):	In the subsumption, it is examined to what extent the conditions (elements) of the definition are given. Here, the facts of the case are weighed against the preconditions from the definitions and the premises (facts). Legal consequences are drawn from the premises, so-called legal claims.
Conclusion	The conclusion is the answer to the major claim. Thus, the case solution reaches a final result here.

Table 1: Core components of the legal writing structure in the appraisal style according to our guidelines.

Structural Components in Legal Case Solutions A persuasive case solution in the appraisal style consists of four main components (see Table 1). The appraisal style always starts with a *major claim*. The *major claim* raises a question, explains the elements of the offense that are to be fulfilled, and is to be written in the subjunctive. *Definitions* define the elements to be fulfilled. The elements always depend on the question raised in the major claim. Only essential passages of the law should be mentioned here; therefore, irrelevant passages should not be annotated. In the *subsumption*, we examine to what extent the conditions (elements) of the definition are given. Here, the facts of the case are weighed argumentatively. This weighing follows established models in argumentation theory (Toulmin et al., 1984; Freeman, 2001). Thus, an argument comprises various elements, including a legal claim and at least one premise that either supports or challenges it. The purpose of the premise is to support the validity of a claim within the context of law by presenting factual statements, legal judgments, or the prevailing opinions of legal experts. It serves as a justification that makes the legal claim understandable. The *conclusion* is the answer to the question that was raised in the major claim. Thus, the case solution here comes to a final conclusion. The question formulated in the major claim is answered. A conclusion is always written in the indicative. Reasons are out of place here; they only belong in the definition or subsumption.

Relations in Legal Case Solutions Apart from the various components that make up the structure of a legal argument, there exist two essential connections between these components. The first pivotal link revolves around the dependence on the major claim and the subsequent conclusion. Every question or issue presented within the major claim must be addressed and resolved in the conclusion. This connection ensures that the arguments presented in support or refutation of the major claim ultimately lead to a clear and definitive conclusion. In other words, the conclusion should provide a resolution to the questions raised in the major claim, tying together the various premises and evidence presented throughout the argument. These relation is illustrated in Figure 2. The subsumption contains the second crucial connection. Here the argumentative elements legal claim and premise are weighed argumentatively. Premises are facts that lead to certain conclusions by the previously attached definition. As a result, the premises back up the legal claims made. More complex combinations of conclusions and premises are feasible. Several different premises can support a legal claim. Equally, a legal claim can be supported by only one premise (see Figure 4 in the Appendix A). A premise might be formed from the facts of the case, past decisions, or the so-called majority view (the majority of legal scholars support a certain interpretation of a fact).

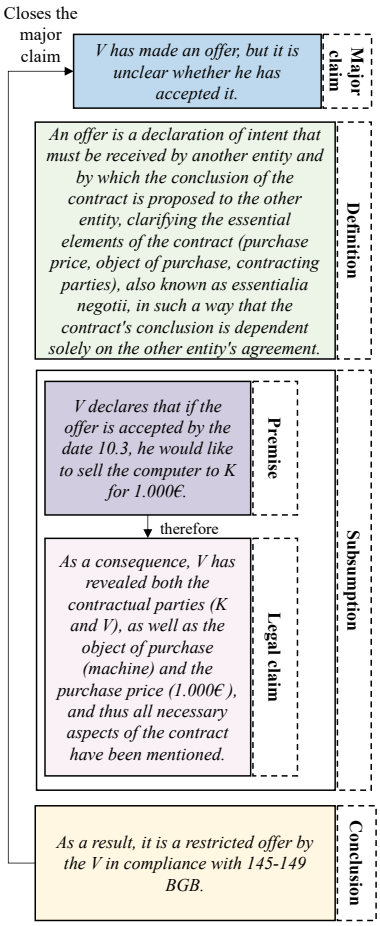

Figure 2: An example of an annotated section from a case solution. We translated the example from German to English for the sake of this paper.

Since it is necessary, especially in jurisprudence, to support the findings obtained in subsumption in a comprehensible way, arguments are used in case solutions in the subsumption to convincingly support the legal conclusion drawn.

3.3 Annotation Process

Two native German speakers independently annotated the legal case solutions according to the components - *major claim, definition, subsumption* (premise and legal claim) and *conclusion* - , as well as for the argumentative relations according to the annotation guidelines we provided. The annotators were trained and educated in the legal domain. Our guidelines consist of thirteen pages. In the guideline[2], we precisely explain and define the components of legal argumentation, which scheme to use for annotation, and how to annotate the subsumption and its argumentative structures (Tettinger, 1982; Backer, 2009; Sieckmann, 2020). In addition, the guideline specifies that sentences failing to meet the criteria of expert opinion style should not be annotated. This decision is based on quality assurance considerations, considering that there may be variations in the texts within the dataset, and not all sentences are expected to align with the requirements of legal writing. Six team workshops were conducted with the annotators as well as a senior researcher to develop a common understanding of the annotation guidelines and to resolve potential disagreements. The senior researcher also has a background in law education, so he was able to assist with legal problems and issues as well. For annotation, we used the tool tagtog[3]. The tool offers the advantages of a graphical interface for marking up units of text and allows monitoring of Inter-Annotator Agreements (IAA) through a dashboard of metrics. Furthermore, the tool has already been used successfully in similar projects (Wambsganss and Niklaus, 2022; Wambsganss et al., 2021). In three workshops, we analyzed the metrics inform of IAAs (e. g., percentage agreement, Kripp. α, Fleiss' Kappa see in Section 4) at 10, 30, and 70 case milestones and highlighted potential difficulties and errors in the guidelines. After annotating the first 100 texts (these texts were each individually edited by both annotators), the two annotators individually annotated the remaining 313 texts. Accordingly, each annotator still annotated 157 texts (or 156 texts) individually. All conflicts in the annotation process were discussed and resolved with three senior researcher. The annotation process was continued on the basis of the agreement; if the agreement was too weak, it was discussed how annotation could be improved. In order to achieve consistency in the annotation process, certain annotation steps needed to be repeated based on this foundation. This was necessary to ensure that the dataset received uniform and accurate annotations throughout. By revisiting these annotation steps, any inconsistencies or discrepancies in the data set could be identified and addressed, allowing for a more reliable and cohesive annotation process. This iterative approach aimed to enhance the overall quality and reliability of the annotations, ultimately leading to a more consistent dataset. Figure 2 shows an example of an annotated part of a case solution with the corresponding components.

[2]The annotation guidelines as well as the entire corpus can be accessed at https://github.com/FlorianRKF-Weber/-structured-and-persuasive-case-solutions-from-law-students.

[3]https://tagtog.net

4 Corpus Analysis

4.1 Inter-Annotator Agreement

To evaluate the reliability of our annotated components and their relationships to each other, we followed the approaches of Stab and Gurevych (2014) as well as Wambsganss and Niklaus (2022) to calculate three different Inter-Annotator Agreements (IAA).

Components	Percentage	Kripp. α	Fleiss' Kappa
Major claim	0.9845	0.9292	0.9292
Definition	0.9720	0.7878	0.7878
Subsumption	0.9622	0.6260	0.6259
Premise	0.9341	0.5590	0.5589
Legal claim	0.9560	0.4502	0.4502
Conclusion	0.9752	0.8836	0.8836
None	0.9026	0.8052	0.8432

Table 2: Inter-annotator agreement of legal component annotations.

Structural Components in Legal Case Solutions To annotate the components of a legal case solution, the annotators determined the individual components in sentences. If a sentence contains a component, it receives one of the labels. Otherwise, it receives the label *none*. Basically, a label can only be assigned to one sentence at a time. An exception is the label of the subsumption. The subsumption is the superior component of the legal claim and premise components. Accordingly, claims and premises must always be subsumptions, but not every subsumption must be a claim or a premise. To represent this circumstance, we have decided to use three models (see Table 5). To evaluate the agreement between the annotators, we compute the percentage agreement p as well as the measures Krippendorff's α (Krippendorff, 1980) and Fleiss' Kappa (Fleiss, 1971). Table 2 illustrates the final resulting IAA values after 100 annotated case solutions. The percentage agreement divides the number of agreements by the label count (Meyer et al., 2014). In order to evaluate the accuracy and reliability of the annotation, a thorough analysis was conducted on the individual values. The results revealed a high level of agreement for the major claim component, with a score of 0.9292, indicating a perfect agreement among the annotators. Similarly, the conclusion component demonstrated a perfect agreement level of 0.8836. However, when examining the premise, legal claim, and subsumption components, the agreement levels were relatively lower, falling below 0.67. Although these components exhibited moderate agreement, they still provided valuable insights for further refinement and clarification. On the other hand, the component related to definition displayed a substantial level of agreement, with a score of 0.7878 according to the (Landis and Koch, 1977). This signifies a significant level of consistency and concurrence among the annotators regarding the definitions within the annotation process. By thoroughly analyzing these individual values, the assessment provided a comprehensive understanding of the effectiveness and reliability of the annotation, highlighting areas of strong agreement as well as identifying aspects that may require further attention and improvement. The evaluation of Fleiss' Kappa comes to similar results. With a total agreement of 0.7751 (Krippendorff's α) and 0.7751 (Fleiss' Kappa), we draw the conclusion that it is consistently possible to annotate argumentative elements in student case solutions. The total agreements show, according to Landis and Koch (1977), a substantial agreement for Fleiss' Kappa and an acceptable agreement for Krippendorff's α (Batanović et al., 2020).

Relations in Legal Case Solutions To assess the reliability of relations, we examined all relations that were annotated in the dataset, i.e., all pairs of a major claim and a conclusion, as well as all pairs of a legal claim and a premise. In total, the markable elements include 3276 pairs, of which 1430 are annotated as legal claims and premises relations, while 2890 of the pairs are annotated as major claims and conclusion relations. We obtained an percentage IAA of 79.7% for the relations between the major claims and the conclusions. The percentage IAA between the claims and premises is 56% (Meyer et al., 2014). Due to this, we calculate the values for Krippendorff's α and Fleiss' Kappa (see Table 6 in the Appendix A). For the relation of major claims and conclusion, we obtained a substantial agreement (0.7750) for Fleiss' kappa (Landis and Koch, 1977) and an acceptable agreement (0.7813) for Krippendorff's α (Krippendorff, 2011; Batanović et al., 2020; Krippendorff, 1980). The relationship between the legal claims and premises shows a fair agreement (0.3979) for Fleiss' kappa (Landis and Koch, 1977). We conclude that component relations and argumentative relations can be reliably annotated in legal case solutions. Nevertheless, it should be noted that the relational agreement between legal claims and

premises according to Krippendorff (2011, 1980) is not acceptable. For the legal claim premise agreement, Fleis' kappa and the percentage agreement, however, indicate acceptable values.

4.1.1 Corpus Statistics

The final corpus comprises 413 case solutions written by students, covering four distinct case studies from the field of civil law. The case solutions are made up of 22,743 sentences totaling and 328,543 tokens (see Table 3). On average, each document has 55.07 sentences and 331.35 tokens. The distribution of the components can be taken from Table 4. Other text fragments were detected as a component with no parameters ("None").

	total	mean	SD	min	max
Sentences	22,743	56.96	27.90	3	133
Tokens	328,543	676.16	331.35	32	1790

Table 3: Overview of the distribution of sentences and tokens in the final corpus. Mean, standard deviation (SD), min and max of sentences and tokens are indicated per document.

	total	mean	SD	min	max
Major claim	3514	8.51	4.76	1	24
Definition	2288	5.54	2.96	1	17
Subsumption	2837	6.87	3.55	1	17
Premise	3304	8.00	4.77	1	27
Legal claim	1949	4.72	2.79	1	17
Conclusion	3531	8.55	4.37	1	23

Table 4: Overview of the distribution of components in the final corpus. Mean, standard deviation (SD), min and max of the annotated components are indicated per document.

5 Application of the Corpus

Modelling Components and Relations of Legal Case Solutions After constructing and analyzing our corpus, we leveraged the novel data to train different ML-models. The detection of the components and relations of legal case solutions is a multi-class classification task. The first task is to classify the single components of the appraisal style. Each sentence can be either a *major claim*, a *definition*, a *subsumption*, a *conclusion* or *non-component*. The second task is the classification of sentences that refer to the component *subsumption*. Each sentence that is a *subsumption* can be a *legal claim*, a *premise*, or a *none* within the subsumption. The third task is to classify the relations between the *legal claims* and the *premises* in a *subsumption*[4]. To perform the classifications, we trained three different text models, we used BERT, RoBERTa, DistilBERT and DistilRoBERTa (see Table 7 in the Appendix A). 20% of the original dataset was used for evaluation, and the remaining 80% for training all the models. The BERT models performed with the highest accuracy, according to our analysis of the models (see Table 7). Therefore, we decided to use three BERT models for the classification. More information about the three models classifier can be found in Table 5 and more information about the performance per class (precision, recall and F1 score) can be found in Table 7. The pre-trained BERT model was acquired from HuggingFace (Wolf et al., 2020) and was subsequently trained using the training dataset. BERT can apply the knowledge it has gained from the initial dataset to the field of legal texts. To make the corpus suitable for sentence-based inputs, it was preprocessed using Spacy[5]. To train the model, we employed 8-piece batches with a maximum sequence length of 128. The BERT models used a warm-up ratio of 0.06, a learning rate of $4e^{5}$, and an Adam epsilon of $1e^{-8}$. For consistency and effectiveness, we adopted the hyperparameters from the pre-trained bert-base-german-cased model and the default parameters of the widely used SimpleTransformers Python library, which have proven successful in similar NLP tasks (Reimers and Gurevych, 2019). Through extensive experiments, we determined that our models performed adequately with the default parameters. We acknowledge the significance of hyperparameter selection and firmly believe that our approach was effective for our specific task and dataset, as demonstrated by competitive results when compared to state-of-the-art models (Wambsganss and Niklaus, 2022).

Extension of the Models by Syntactic Rules To better meet the prediction requirements of a legal case solution, we have extended the model with syntactic rules. In the first step, we have added the identification of headings to the model. Thus, all sentences that begin with a Roman or Arabic numeral, e.g., 1, II, or with a letter (a., a), A), etc.) are marked as headings. This is important because the

[4]The model does not incorporate the relations between the major claim and the conclusion, as we determined that simple heuristics offered a superior approach to providing feedback to students regarding the connection between these two components.

[5]https://spacy.io

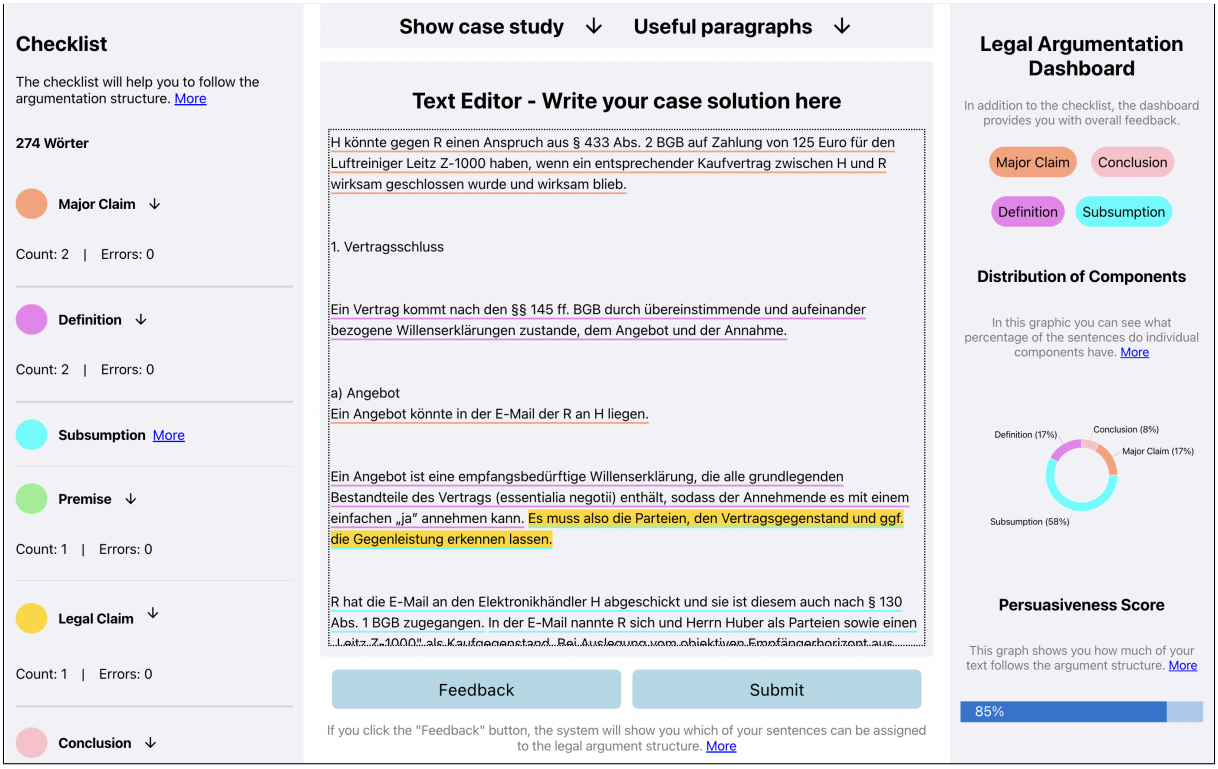

Figure 3: Screenshot of our writing support system with the integration of the three BERT-models. To facilitate understanding we translated the user interface to English.

Model	Description
1. Main model (BERT) - multi-class prediction model	Identifying the fundamental elements of the appraisal style (major claim, definition, subsumption and conclusion).
2. Secondary model (BERT) - multi-class prediction model	Identifying the components of an argument in the subsumption (legal claims and premises).
3. Tertiary model (BERT) - binary classification model	Identifying relationships between the components in the subsumption (claims and premises).

Table 5: Overview and description of the three BERT models and its classifier. We show the main model, the secondary model and the tertiary model.

headings have nothing to do with the specified components but are often used by students to structure their case solutions (see Figure 3). In the second step, we have defined a collection of abbreviations typically used in legal case solutions (see Table 8 in the Appendix A). These abbreviations played a vital role in our sentence-based models by ensuring that sentences would not be inappropriately segmented due to the presence of punctuation marks immediately following the abbreviations. By incorporating these abbreviations into the models, we maintained the integrity and coherence of the text, enabling more accurate and effective analysis of the legal case solutions.

Writing Support System for Persuasive Legal Case Solutions We have designed a writing support system in which we have implemented the three BERT models as feedback algorithms. The system is based on a user-centered design and fundamentally follows the theory of learning from errors (Metcalfe, 2017) and supports learners through scaffolding (Wong and Lim, 2019; Cagiltay, 2006). Based on our three models, the system can provide individual feedback to students in German law courses during their writing process (Hattie and Timperley, 2007). Our writing support system is presented in Figure 3.

Evaluation in a Writing Task We evaluated our writing support system in an online experiment with 34 students who were enrolled in a law or business law program. We selected Prolific[6] as the experimental platform due to its consistent track record of delivering high response quality and a diverse range of samples, making it one of the most reliable platforms for conducting behavioral research (Peer et al., 2017). In the online exper-

[6] https://www.prolific.co

iment, students were asked to solve a legal problem with the assistance of our writing support system. The students were randomly divided into two groups. The control group solved the problem with a reduced version of our system, receiving only static feedback on persuasive writing of case solutions. The treatment group solved the task with feedback based on the three BERT models. Apart from the feedback, the two versions of the system were identical in design to maintain consistency. Both systems use a case study, useful paragraphs, and a checklist (see Figure 3). Before the students started the writing task, we conducted a pre-test to make sure that they had the same knowledge about legal writing. In the pre-test, students received two predefined case solutions and were required to rate them on a scale of 1–5. The scale indicates how qualitatively well the case solutions were written (see Table 9 in the Appendix A).

After the interaction with the system, the participants were asked to complete a post-survey to measure the learning outcome between the two versions of the systems. In the post-test, students were again asked to rate two predefined cases (scale of 1–5) and explain why they rated the cases accordingly (assessment task). After the post-test, students were asked questions regarding their perception of the system (post-survey) (see Table 10 in the Appendix A). To evaluate the perception of the system, the students were asked questions about the technology acceptances of the system (Venkatesh and Bala, 2008) and the feedback accuracy of the system (Podsakoff and Farh, 1989). To test whether participants conscientiously completed the surveys, we included two control questions.

Results After analyzing that all participants were either law students or had already attended a law lecture in the course of their study program, had sufficient knowledge of German, and could answer both control questions, we obtained 29 valid responses (14 treatment group, 15 control group). The participants had an average age of 25.83 (SD = 5.66). Among them, there were 9 females, 16 males, and four individuals who identified as non-binary. The participants spent between 30 and 55 minutes writing the case solution. The post-test required approximately 20 minutes to complete. To assess participant responses, we employed a standardized 7-point Likert Scale commonly used in psychology. In this scale, the value of 4 represents neutrality. Values higher than 4 indicate positive outcomes and provide evidence of the system's effective design. The perceived usefulness has a value of 5.07 (SD = 1.14). Perceived usefulness (PU) shows if the users believe in an increased value by using the system (Davis, 1989). Perceived ease of use (PEOU) was also rated by participants above the neutral value of 4 (mean = 5.5, SD = 1.08). PEOU promotes intrinsic learner motivation and can lead to increased learning success (Barto et al., 2004). Finally, the participants also rated the intention to use (ITU) with a mean value of 5.43 (SD = 0.99), which is above the neutral value. The ITU indicates that the participants would use the corresponding system in a law course (Agarwal and Karahanna, 2000). We also analyzed feedback accuracy to determine whether participants perceive the feedback algorithm to be accurate (Podsakoff and Farh, 1989). The results show that participants rate the feedback with a mean value of 4.95 (SD = 0.74) hence 0.95 higher than the neutral value 4.

In addition to participants' perceptions of the system, we also measured the learning success of the system in a post-test. In the post-survey, we indicated that the participants performed significantly better than the control group at the assessment task (p-value = 0.0404, W = 146.5)[7]. At the same time, we could show that there were no significant differences between the two groups in the pre-test, which excludes a bias by a control group with possibly more knowledge (see Table 11 in the Appendix A).

6 Conclusion

In this research, we offer a novel scheme for annotating structured elements and arguments in student-written case solutions. We used the scheme to create a corpus of 413 students' written legal case solutions, which consists of 25,103 sentences and 310,363 words. Furthermore, we present an annotation study based on 100 case solutions and show that the annotation of student-written case solutions is possible. Finally, we integrated and evaluated three trained BERT-models based on our corpus in a writing support system. In order to improve teaching in large-scale learning settings, we expect that integrating the provided annotation scheme and our argumentation corpus would encourage the creation of writing support systems.

[7]We performed a two-tailed Wilcoxon rank-sum test.

Limitations

Regarding our work, a few limitations should be mentioned. During the annotation process, conflicts between annotations occurred. All conflicts were discussed with a senior researcher and resolved in this way. Thus, we reached the best possible agreements, but still some agreements are lower than others (see Table 2). For example, legal claims and premises have a relatively large room for interpretation. Perfect results can only be expected by over-anchoring the annotators and weakening the guideline, which we have consciously avoided in our research. The comparison of the IAA with other works in the field of NLP from legal science is not possible, because the works either do not examine the components of the appraisal style or identify the components of the judgment style without the indication of the IAAs (Urchs et al., 2020). Compared to works that also annotated premises (Kripp. α = 51.08%) and claims (Kripp. α = 55.49%) in business pitches (Wambsganss and Niklaus, 2022), our work provides comparable results with respect to the agreement of the Krippendorff α (premise = 55.89%, legal claim = 45.02%). Further work shows similar results α = 44.1% (Park and Cardie, 2018). All in all, we can assume that both our components and our mounted relationships, achieve comparable or better results than comparable works (e.g., Park and Cardie (2018)).

Although our model shows accurate values between 78% and 92% for predicting the components of the appraisal style, the values for determining legal claims and premises are lower (62% and 78%) compared to the other values. However, they display reasonable values when compared to previous NLP studies. We can only compare our work to other related work in another domain because values for detecting legal claims and premises are not available in the NLP literature. For instance Wambsganss and Niklaus (2022), present an accuracy of 54.12% for their Long Short-Term Memory (LSTM) model which detects claims and premises. With the mentioned model the authors shows positive outcomes in supporting students' argumentative skills. Our models show similar or higher precision in comparison to the works of Poudyal et al. (2020) or Wambsganss and Niklaus (2022) (see Table 7 in the Appendix A) and our post-test results also show significant learning outcomes (see Table 11 in the Appendix A). Although we can show a significant learning output, it must be noted that this is only short term. As a result, we intend to carry out additional field experiments in the future to establish the system's effectiveness over a more extended period and demonstrate long-term success.

As a third possible limitation, our models are limited to applying the appraisal style in German only. In the future, further efforts have to be made to investigate the transfer-ability or adaptation of our models to other countries with other legal systems and other languages. However, we assume that this is possible in principle, since some countries such as China now use the appraisal style in law teaching (Man, 2022) and countries such as the U.S. use at least similar approaches such as learning with case studies using the IRAC formula (Metzler, 2002). Nevertheless, some adaptation of the models is needed, since the language and the legal form in each country have their own specificities.

Ethics consideration

It is important to acknowledge that this research was conducted by a diverse team of authors and annotators with backgrounds encompassing Western European, Asian, female, and male perspectives.

All data collection procedures strictly adhered to the ethical and privacy policies outlined by our university and the respective platforms involved. Prior to participation in surveys or interviews, all participants were duly informed about the data processing procedures and provided their explicit consent. To ensure privacy, all data were anonymized during analysis and could be deleted upon the request of participants.

In collaboration with our university, we conducted a comprehensive risk assessment and ethics review for this project. The findings from both investigations affirm that the project does not pose any risks to the students. Our models and the system utilizing them do not present any potential dependencies or hazards that could negatively impact students. It is worth noting that similar models have been trained in the past, aimed at enhancing students' argumentation skills among other objectives. Based on our current knowledge, no risks have been identified associated with the utilization of these models.

We are committed to upholding the highest ethical standards throughout our research, prioritizing the well-being of all participants involved.

References

Ritu Agarwal and Elena Karahanna. 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, 24(4):665–694.

Carsten Backer. 2009. Der Syllogismus als Grundstruktur des Juristischen Begrundens. *Rechtstheorie*, 40(3):404–424.

Andrew G Barto, Satinder Singh, Nuttapong Chentanez, et al. 2004. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of international conference on developmental learning*, pages 112–119.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2020. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLoS One*, 15(11):1–30.

Michael Beurskens. 2016. Neue Spielräume durch Digitalisierung? E-Learning in der deutschen Rechtslehre. *ZDRW Zeitschrift für Didaktik der Rechtswissenschaft*, 3(1):1–17.

Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, 21(1):5–31.

Kursat Cagiltay. 2006. Scaffolding strategies in electronic performance support systems: Types and challenges. *Innovations in education and Teaching International*, 43(1):93–103.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 13(3):319–340.

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.

Cecilie Enqvist-Jensen, Monika Nerland, and Ingvill Rasmussen. 2017. Maintaining doubt to keep problems open for exploration: An analysis of law students' collaborative work with case assignments. *Learning, culture and social interaction*, 13:38–49.

Lisa K Fazio and Elizabeth J Marsh. 2009. Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, 16(1):88–92.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

James B Freeman. 2001. Argument structure and disciplinary perspective. *Argumentation*, 15(4):397–423.

Isabelle Gauer, Hanjo Hamann, and Friedemann Vogel. 2016. Das juristische Referenzkorpus (JuReko)-Computergestützte Rechtslinguistik als empirischer Beitrag zu Gesetzgebung und Justiz. In *DHd 2016: Modellierung - Vernetzung - Visualisierung*, page 129–131.

Ben Hachey and Claire Grover. 2005. Automatic legal text summarisation: experiments with summary structuring. In *Proceedings of the 10th International Conference on Artificial intelligence and Law*, pages 75–84.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Bethany Rubin Henderson. 2003. Asking the lost question: what is the purpose of law school. *Journal of Legal Education*, 53(1):48–79.

Constantin Houy, Tim Niesen, Peter Fettke, and Peter Loos. 2013. Towards automated identification and analysis of argumentation structures in the decision corpus of the german federal constitutional court. In *2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, pages 72–77. IEEE.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to Give Feedback: Modeling Attributes Affecting Argument Persuasiveness in Student Essays. In *IJCAI*, pages 4130–4136.

Susan Hanley Kosse and David T Butle Ritchie. 2003. How judges, practitioners, and legal writing teachers assess the writing skills of new law graduates: A comparative study. *Journal of Legal Education*, 53(1):80–102.

Klaus Krippendorff. 1980. *Validity in content analysis*. Campus, New York.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Departmental Papers (ASC)*, pages 1–10.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *International Biometric Society*, pages 159–174.

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

Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102–157.

JIN Man. 2022. The Appraisal-Based Case Teaching Method in China's Legal education. *Canadian Social Science*, 18(2):1–4.

Janet Metcalfe. 2017. Learning from errors. *Annual Review of Psychology*, 68:465–489.

Jeffrey Metzler. 2002. The importance of IRAC and legal writing. *University of Detroit Mercy Law Review*, 80:501–514.

Christian M Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. 2014. DKPro agreement: An open-source Java library for measuring inter-rater agreement. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: System demonstrations*, pages 105–109.

Raquel Mochales and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments? A case study on the legal argumentation of the ECHR. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 21–30.

Raquel Mochales and Marie-Francine Moens. 2008. Study on the Structure of Argumentation in Case Law. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 11–20, Amsterdam, The Netherlands, The Netherlands. IOS Press.

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

Raquel Mochales-Palau and M Moens. 2007. Study on sentence relations in the automatic detection of argumentation in legal cases. *Frontiers in Artificial Intelligence and Applications*, 165:89–99.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Hans-Joachim Musielak and Wolfgang Hau. 2005. *Grundkurs BGB*, 17 edition. CH Beck, München.

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163.

Philip M Podsakoff and Jiing-Lih Farh. 1989. Effects of feedback sign and credibility on goal setting and task performance. *Organizational behavior and human decision processes*, 44(1):45–67.

Prakash Poudyal, Teresa Gonalves, and Paulo Quaresma. 2019. Using Clustering Techniques to Identify Arguments in Legal Documents. In *Third Workshop on Automated Semantic Analysis of Information in Legal*, pages 1–8.

Prakash Poudyal, Jaromír Šavelka, 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.

Chris Reed. 2006. Preliminary results from an argument corpus. *Linguistics in the twenty-first century*, pages 185–196.

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*, pages 3982–3992, Hong Kong, China.

Jan-R Sieckmann. 2020. *Logik juristischer Argumentation*, 76 edition. Nomos Verlag, Baden-Baden.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990.

Carl-Friedrich Stuckenberg. 2020. Der juristische Gutachtenstil als cartesische Methode. *ZDRW Zeitschrift für Didaktik der Rechtswissenschaft*, 6(4):323–341.

Peter Josef Tettinger. 1982. *Einführung in die juristische Arbeitstechnik*, 81 edition. Beck, München.

Stephen Toulmin, Richard D. Rieke, and Allan Janik. 1984. *An Introduction to Reasoning*, 2 edition. Macmillan.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2020. Towards Classifying Parts of German Legal Writing Styles in German Legal Judgments. In *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 451–454. IEEE.

Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences*, 39(2):273–315.

Vern Walker, Karina Vazirova, and Cass Sanford. 2014. Annotating patterns of reasoning about medical theories of causation in vaccine cases: toward a type system for arguments. In *Proceedings of the first workshop on argumentation mining*, pages 1–10.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling Persuasive Discourse to Adaptively Support Students' Argumentative Writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020a. Al: an adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020b. A corpus for argumentative writing support in german. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 856–869.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. *In 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*, pages 4063–4077.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

Sarah Shi Hui Wong and Stephen Wee Hun Lim. 2019. Prevention-permission-promotion: A review of approaches to errors in learning. *Educational Psychologist*, 54(1):1–19.

A Appendix

Relations	Percentage	Kripp. α	Fleiss' Kappa
Major claim and conclusion	0.7970	0.7813	0.7750
Legal claim and premise	0.5600	0.4147	0.3979

Table 6: Inter-annotator agreement of relations annotations.

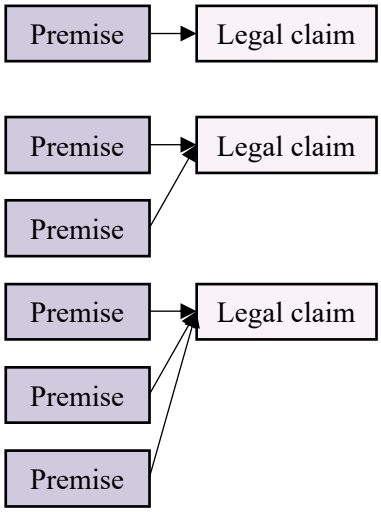

Figure 4: The figure demonstrate various support linkages between premises and legal claims. One or more premises can be used to support any legal claim.

Classifier Model Name	Model Type	Precision	Recall	F1 Score	Class
Lawful Components	BERT	0.92	0.95	0.93	MC
		0.87	0.92	0.89	C
		0.78	0.86	0.82	D
		0.69	0.73	0.71	S
		0.91	0.86	0.88	N
	RoBERTa	0.93	0.96	0.94	MC
		0.90	0.90	0.90	C
		0.79	0.89	0.84	D
		0.60	0.80	0.69	S
		0.93	0.84	0.88	N
	DistilBERT	0.95	0.95	0.95	MC
		0.87	0.91	0.89	C
		0.82	0.86	0.84	D
		0.62	0.71	0.66	S
		0.92	0.87	0.89	N
	DistilRoBERTa	0.91	0.93	0.92	MC
		0.81	0.88	0.84	C
		0.80	0.77	0.78	D
		0.53	0.67	0.59	S
		0.89	0.82	0.86	N
Subsumption Types	BERT	0.78	0.58	0.66	LC
		0.62	0.79	0.69	P
		0.83	0.74	0.78	N
	RoBERTa	0.76	0.51	0.61	LC
		0.58	0.68	0.63	P
		0.77	0.76	0.76	N
	DistilBERT	0.67	0.57	0.62	LC
		0.61	0.72	0.66	P
		0.79	0.73	0.76	N
	DistilRoBERTa	0.56	0.51	0.53	LC
		0.52	0.66	0.58	P
		0.77	0.63	0.69	N
Claim-Premise Relation	BERT	0.90	0.90	0.90	-
	RoBERTa	0.89	0.60	0.72	-
	DistilBERT	0.88	0.87	0.88	-
	DistilRoBERTa	0.74	0.98	0.85	-

Table 7: Comparison of the different types of models. Each model type was applied to the different classifiers. MC = Major Claim, C = Conclusion, D = Definition, S = Subsumption, LC = Legal Claim, P = Premise, and N = None.

Abbreviation (German)	**ff.**	**abs.**	**art.**	**gem.**	**nr.**	**ggf.**
Translation	et seqq.	para.	art.	acc. to	no.	-
Meaning	Refers to further paragraphs.	paragraph	article	according to	number	if applicable
Abbreviation (German)	**abl.**	**abschn.**	**abschl.**	**allg.**	**anm.**	**ausf.**
Translation	-	sec.	-	-	-	-
Meaning	deprecating	section	markdown	in general	comment	in detail
Abbreviation (German)	**vgl.**	**i.S.d.**	**insbes.**	**grds.**	**ggü.**	**bzw.**
Translation	-	-	-	-	-	resp.
Meaning	see	in the sense of	notably	in principle	vis-à-vis	respectively
Abbreviation (German)	**bzgl.**	**bspw.**	**bsp.**	**betr.**	**begr.**	**Beschl.**
Translation	-	e.g.	-	-	-	-
Meaning	regarding	for example	example	concerning	justifying	resolution

Table 8: Overview of abbreviations with which the models were extended. The models understand the appropriate abbreviations as such and do not break up sentences. The list will be extended in the future.

Section	Variables	Items	Scale
Pre-Survey	Previous experience with legal writing	Have you already taken or completed a law class? (*This also includes courses such as introduction to law or similar courses that are offered, for example, as part of a business administration degree program.*)	Yes / No
Pre-Survey	Previous experience with legal writing	In which field are you studying or have you studied ?	Law, Business Law, Business Administration, Business Sciences
Pre-Survey	Demographics	1. Age 2. Gender 3. Language	Open
Pre-Test	Checking the level of knowledge before interacting with the system	Evaluation case solution 1.1 (civil law - rather good solution) Evaluation case solution 1.2 (civil law - weak solution)	Scale 1-5 (good, rather good, average, rather weak, weak) + Open (Open question for the explanation of the evaluation)
Writing Task	Online assignment	"In the following, you can solve the civil law case. Use the appraisal style. The writing support system will help you write your case solution and will also provide you with the exact facts of the case in the form of a case study. Your case solution should be about 350-450 words (the system will show you your word count)."	Open question

Table 9: Overview of the pre-survey, pre-test and the writing task.

Section	Variables	Items	Scale
Post-Test	Checking the level of knowledge after interacting with the system (assessment task)	Evaluation case solution 2.1 (civil law - weak solution) Evaluation case solution 2.2 (civil law - good solution)	Scale 1-5 (good, rather good, average, rather weak, weak) + Open (Open question for the explanation of the evaluation)
Post-Survey	Intention to use (Agarwal and Karahanna, 2000)	"Assuming the system would be available for a law course, I would use it again." "Assuming the system would be available at a law course, I would plan to use it."	1- 7 Likert scale (7: highest)
Post-Survey	Perceived usefulness (Agarwal and Karahanna, 2000)	"Using the writing support system helps me more effectively write persuasive case solutions using the appraisal style." "I find the interaction with the system useful in writing persuasive case solutions using the appraisal style."	1- 7 Likert scale (7: highest)
Post-Survey	Perceived ease of use (Venkatesh and Bala, 2008)	"Learning how to use the system would be easy for me." "I perceived the interaction with the system as easy." "I think it would be easy for me to become skillful in using the system."	1- 7 Likert scale (7: highest)
Post-Survey	Feedback accuracy (Podsakoff and Farh, 1989)	"The systems evaluation of my case solution reflects my actual performance." "The systems has accurately evaluated my performance." "The recommendations I received from the system was an accurate assessment of my performance." "I assume that the system will help me improve my ability to write persuasive case solutions in the appraisal style."	1- 7 Likert scale (7: highest)
Post-survey	Control questions	"Please check "Strongly agree." "A certain word was mentioned in the system tutorial video. Please write this word in the text box below."	Open question

Table 10: Overview of the post-survey and the post-test.

Group	p-value	W	Mean (TG)	Mean (CG)	SD (TG)	SD (CG)
Pre-Test	0.638	115	0.286	0.233	0.323	0.319
Post-Test	0.0404	146.5	0.357*	0.133	0.305	0.229

Table 11: Results of the analysis of the *learning outcome*. We show the mean, the standard derivation, Wilcoxon statistic (W), middle rank of the control group and the treatment group, as well as the results of a Wilcoxon rank-sum test. We set the significance level at alpha 0.05: p<=0.05*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 (Introduction).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 5 (Application of the Corpus).

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. We have developed the system ourselves and there is still no cited source in which the system appears.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. We have developed the system ourselves.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. We have developed the system ourselves.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
All the data we have collected has been only processed anoymized. In the course of our publication, no connections to personal data can be made.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5 (501-513).

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Training of the models for the system. See Section 5 (442-483).

C ☑ Did you run computational experiments?

Section 5 (442-483).

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No, because we can't report anything in the paper. However, we can show the data here. Infrastructure for training: M1 Pro (Apple Silicon) Time for fine-tuning: The majorclaim/subsumption/conclusion/definition/none classifier: 50 minutes The premise/claim/none classifier: 15 minutes The relation classifier: 10 minutes

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
For the number of parameters we refer to this work: https://arxiv.org/pdf/1810.04805.pdf. Additionally we 110 million (since we used BERT-base).

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5 (442-483) and the Appendix.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
See Guideline in the supplementary material.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 5 (514-556).

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 8

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We made a request to the Ethics Committee of our university. This was accepted.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.3 Section 8