Argumentation Structure Prediction in CJEU Decisions on Fiscal State Aid

Piera Santin*
Alma AI, University of Bologna
Italy

Giulia Grundler*
DISI, University of Bologna
Italy

Andrea Galassi†
DISI, University of Bologna
Italy

Federico Galli
Alma AI, University of Bologna
Italy

Francesca Lagioia†
Alma AI, University of Bologna
European University Institute
Italy

Elena Palmieri
DISI, University of Bologna
Italy

Federico Ruggeri
DISI, University of Bologna
Italy

Giovanni Sartor
Alma AI, University of Bologna
European University Institute
Italy

Paolo Torroni
DISI, University of Bologna
Italy

ABSTRACT

Argument structure prediction aims to identify the relations between arguments or between parts of arguments. It is a crucial task in legal argument mining, where it could help identifying motivations behind judgments or even fallacies or inconsistencies. It is also a very challenging task, which is relatively underdeveloped compared to other argument mining tasks, owing to a number of reasons including a low availability of datasets and a high complexity of the reasoning involved. In this work, we address argumentative link prediction in decisions by Court of Justice of the European Union on fiscal state aid. We study how propositions are combined in higher-level structures and how the relations between propositions can be predicted by NLP models. To this end, we present a novel annotation scheme and use it to extend a dataset from literature with an additional annotation layer. We use our new dataset to run an empirical study, where we compare two architectures and explore different combinations of hyperparameters and training regimes. Our results indicate that an ensemble of residual networks yields the best results.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Computing methodologies** → **Language resources**; **Information extraction**.

KEYWORDS

Argument Mining, Legal Argument, Link Prediction, CJEU decisions

*Both authors contributed equally to this research.
†Corresponding authors.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595174

ACM Reference Format:
Piera Santin, Giulia Grundler, Andrea Galassi, Federico Galli, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2023. Argumentation Structure Prediction in CJEU Decisions on Fiscal State Aid. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594536.3595174

1 INTRODUCTION

Legal argumentation draws on logic, philosophy and linguistics to explore how different statements and opinions are proposed, discussed and assessed, giving an understanding of points being made, the relationships between them and how they support (or undermine) a certain conclusion. Computational legal argumentation, at the intersection of legal argumentation, computer science and artificial intelligence (AI), is one of the most lively research areas of AI and law [5, 6]. Indeed, the legal domain offers a natural setting for the application of argument models as well as machine learning and natural language processing (NLP) methods to (i) automatically retrieve legal arguments from large corpora and predict the relations among their different components [15, 31, 32, 46], (ii) summarize and classify legal texts [18], (iii) perform legal reasoning [2, 3, 38], build ontologies [21] and (iv) support the teaching of law [1]. The development of automated methods for the argumentative analysis of legal documents could have a tremendous impact on many areas of the law, providing valuable instruments to support legal research and facilitate the everyday work of legal practitioners. In this respect, in recent years, there has been significant progress in argument mining (AM), a discipline that combines methods from NLP, machine learning and computational argumentation to address a variety of tasks, usually but not necessarily addressed in a sequence [23]. These include relatively simple tasks such as argument component detection and classification as well as more complex ones such as argument structure prediction. Detecting argument components, such as claims or evidence, focuses on identifying which sentences are argumentative. Sometimes the task is further specialized into discriminating between different types of argumentative content or identifying the boundaries of the argument components within or

across sentences [4, 29, 32]. Argument structure prediction instead amounts to identifying the relations between arguments and/or their components [7, 24]. In other words, it aims to reconstruct the reasoning path that connects argumentative elements scattered in the text at hand [23, 24]. This is a highly challenging task, even for domain experts, as it involves high-level reasoning. First of all, the argument relations may be heterogeneous, like supports and attacks. In the latter case, it is possible to further distinguish between rebuttals (challenging the acceptability of a proposition) and undercuts (challenging the acceptability of an inference between two propositions). Secondly, in natural language texts, the way argumentative propositions and their underlying relations are presented is pretty far from the prototypical argumentation patterns in theoretical studies. This is particularly true in the judicial discourse where complex argument patterns are presented, multiple inferences are merged, concepts are repeated and paraphrased, and premises are left implicit.

The judiciary has been one of the first AM application areas [29]. There, it would ease the identification and classification of (i) arguments in court rulings; (ii) their possible fallacies and inconsistencies; (iii) the similarities and discrepancies among judgments; and (iv) the reasons behind the final outcome of cases. However, despite the significant development of AM methods for other domains such as social media analysis, scientific literature, and politics [26], there is still a lack of development and deployment of AM applications to legal texts, particularly to judicial decisions [53]. This is partly due to the dearth of comprehensive corpora for research. The complexity and labour cost required to produce new annotated corpora certainly constitute a major barrier in this area.

In this study, we contribute to AM research by focusing on one of the most challenging AM tasks, argument structure prediction, which so far has been only partially addressed in the legal domain. To this end, we propose an approach for predicting the relations between propositions – i.e., argument components– in judicial decisions by the Court of Justice of the European Union on Fiscal State Aid. Our paper builds upon previous work focused on the detection and classification of premises and their intermediate outcomes (which are also called premise(s)– since they support further outcomes(s))– or the final conclusion [16]. The focus of this work is a study of how propositions are combined in higher-level structures and how the relations between propositions can be predicted by NLP models. To this end, we enrich the Demosthenes data set introduced in [16] with an additional layer of annotation so as to capture the inferential connections between propositions. Compared to other works, our annotation scheme distinguishes between different types of support and attack relations, also establishing an additional, even though minor, type of connection (see section 3).

The paper is structured as follows. Section 2 provides an overview of related works. Section 3 describes the updated corpus and the annotation procedure. Section 4 concerns the experimental setting. Section 5 presents the results. Section 6 provides conclusions and indicates future developments.

2 RELATED WORK

A crucial obstacle to providing effective automatic support to legal argumentation pertains to the knowledge acquisition bottleneck. Argument mining tasks depend on the availability of high-quality annotated corpora to train and evaluate the performance of automated approaches. Creating these corpora is a labour-intensive, complex and time-consuming process, which requires the guidance of legal experts familiar with the features specifically characterizing legal argumentation. Indeed, there is a discrepancy between computational and legal approaches in analysing, modelling and annotating arguments in court decisions [17]. While NLP researchers often treat arguments as mere structures of premises and claims [44], in some areas of legal research it is often crucial to distinguish among different kinds of arguments, classify them according to the rich typology that is rooted in the theory and practice of legal argumentation, and capture the complex and heterogeneous structure of inference relations [48]. Unfortunately, the sample of annotated corpora meeting these requirements is very limited. As detailed in the following, (i) few works have analysed how natural language argumentation is used in real courts [17, 29, 37, 45]; (ii) they cover a very small number of legal fields and judicial bodies (mostly European Court of Human Rights decisions); (iii) the existing annotations mainly concerns the distinction between argument components (i.e., premises and conclusions). Thus the research community can highly benefit from the availability of new datasets and annotations, covering different areas of the law and attempting to classify the arguments and their relations according to a legal typology.

Since the early stage of argument mining research in the legal domain [30], decisions by the European Court of Human Rights (ECHR) have been used as a practical application scenario for detecting and classifying argument components and predicting inter-argument connections. In their studies, Mochales and Moens [28–30] annotated a corpus of 47 ECHR decisions, differentiating between premises, conclusions, and non-argumentative sentences, by taking inspiration from the argumentation model developed by Walton [49]. Concerning the relation prediction task, it was only briefly touched upon. Indeed, only premise-conclusion relations were studied due to the inability of the used context-free grammar to take into account complex inferential links inside an argument. Again in the context of ECHR judgments, Teruel et al. [45] produced a corpus of 7 ECHR judgments, annotated with (i) argument components – classified as major claims, claims and premises based on the Toulmin theoretical model [47] – and (ii) the connections between them, mainly support and attack relations. More recently, Poudyal et al. [37] published a corpus of 42 ECHR judgments, reproducing the annotation scheme developed by Mochales and Moens [29]. They evaluated three tasks, i.e., (i) clause detection to identify whether or not a clause belongs to an argument; (ii) premise and conclusion recognition; and (iii) argument relation prediction. The latter was designed to assess, for each pair of arguments, whether they are related, though without dealing with more complex structures or distinguishing between different kinds of links. This approach represents, to date, one of the few works whose goal was to implement a full-fledged argumentation mining system specific to a single legal domain. ECHR decisions were also used by Habernal et al. [17], who departed from the usual premise-conclusion scheme by developing a novel annotation scheme to cover the different actors from which the arguments originate (e.g., the ECHR, the Applicant, the State, the Commission), as well as multiple argumentation schemes (e.g., procedural, interpretative, institutional arguments).

Grabmair et al. [15] proposed the Legal UIMA system to extract argument-related semantic information from a set of U.S. Court of Federal Claims cases, deciding whether compensation claims comply with a federal statute establishing the National Vaccine Injury Compensation Program.

With the aim of summarizing court decisions, Yamada et al. [51] annotated 89 Japanese civil case judgments. The task of argument structure extraction was divided into four sub-tasks: (i) *Issue Topic Identification*, to find sentences that describe an Issue topic; (ii) *Rhetorical Status Classification*, to determine the rhetorical status of each sentence; and (iii) *Issue Topic Linking*, to associate each sentence with exactly one Issue Topic; and (iv) *Framing Linking*, to connect two sentences if one provides argumentative support for the other. To this end, the corpus was annotated to classify each sentence into one of the seven defined classes (e.g., fact, background, conclusion). With a similar aim to improve case summaries, Xu et al. [50] explored AM by annotating human-made summaries, according to what they called "legal argument triples", containing the following information: (i) the main issues addressed by the Court; (ii) the court's conclusion on each issue; and (iii) a description of the reasons given by the court to support its conclusion. The annotated corpus covers different kinds of legal claims and issues presented before Canadian courts.

Most related to our study, Grundler et al. [16] recently released a dataset containing 40 decisions by the CJEU on Fiscal State Aid. The annotation specifies three hierarchical levels of information: the argumentative elements, their types (i.e., legal or factual), and their argument schemes (e.g., arguments from precedent, from a rule, from interpretation, from authority).

Turning to methodologies, one of the first approaches for identifying the links between argument components is described by Palau and Moens [33]; their approach is based on the creation of a Context-Free Grammar to parse the argument structure, and they indicate an accuracy of around 60%. Poudyal et al. [36] use the Fuzzy c-means clustering algorithm to group the argumentative components, allowing each sentence to be in more than one cluster; they report a macro F1 of 0.497 and cluster purity of 0.499. Poudyal et al. [37] approached the problem as a sentence pair classification and obtained an F1 score of 0.511, employing RoBERTa with an additional classification layer on top. Similarly, Stab and Gurevych [42] classify each pair of argument components as support or non-support, reaching an F1 score of 0.722 with an SVM model.

Since the different tasks included in argumentation mining are interrelated, some approaches jointly address them, with the idea of transferring knowledge between them. Stab and Gurevych [43] used Integer Linear Programming and SVMs to jointly perform component classification and link prediction on persuasive essays. They reach an F1 score of 0.751 on link prediction and 0.680 on relation prediction, with support and attack types. In the work by Niculae et al. [32], argument classification and link prediction are jointly performed on two corpora with a structured learning framework based on a factor graph. The results vary depending on the dataset, between an F1 score of 50.0 and 68.9. Recently, Sazid and Mercer [40] developed a deep learning architecture and a novel unified representation that combines the component and relation identification into a single sequence tagging problem, which reaches an F1 score of 52.71 in link prediction.

Concerning the task of relation prediction, Cocarascu and Toni [9] used two LSTMs for the identification and classification of attack, support or neither relations. They reach an F1 score of 89.07, but with a balanced dataset with only 37% of relations not having any link. Cheng et al. [8] propose a multi-task learning framework, based on hierarchical bidirectional LSTMs with a conditional random field, for detecting arguments and linking argument pairs of sentences in reviews and rebuttals. In [41], given a quotation and five candidate replies with their context, the model decides which reply is related to the quotation, performing a binary classification task. A similar task is applied to the legal field in [52], where the participants of the Argmine Challenge had to identify the correct defence argument linked to the given plaintiff argument among five candidates.

All these efforts lack a shared argumentative model, which makes the results and the problem addressed not directly comparable. At the same time, most of the literature on argument structure prediction stresses the complexity of the problems addressed, due to the diversity of domains, the scarcity of data, and unbalanced datasets.

3 DATASET

In this section, we describe the source corpus and the methodology for its annotation. The final corpus is publicly available. [1]

3.1 The source corpus

The analysed documents were retrieved from the Demosthenes corpus, recently published by Grundler et al. [16]. Demosthenes consists of 40 decisions on fiscal state aid by the Court of Justice of the European Union (CJEU), ranging from 2000 to 2018. Its annotations distinguish between two argumentative elements, i.e., premises and conclusions, as part of an argument chain. An argument chain is defined as an argument supporting the final conclusion concerning a specific ground of appeal, together with all counterarguments considered by the Court. More than one argument chain may be provided in a single decision. The argumentative elements are characterized by a set of attributes and their possible values. In particular, each premise and conclusion is denoted through a unique identifier (`ID`), whose value is constructed by joining a letter (which denotes the argument chain to which the premise or conclusion belongs, e.g. A or B) with a progressive number (which indicates the single premise or conclusion within the chain, e.g., A1, A2, An; B1, B2, Bn). Premises are also distinguished in *factual* and *legal*. Table 1 reports the pre-existing annotation scheme.

All documents are in English. We chose this source since: (a) CJEU decisions contain a rich and diverse set of legal arguments characterized by different kinds of inferential connections (e.g., support, rebuttal, undercut); (b) they have a standard (although not fixed) structure, in which argument chains are embedded and can be easily identified; (c) the selected decisions come from the same domain, i.e., fiscal State aids, which strongly relies on judicial interpretation; and (d) our annotators have some expertise in this domain. CJEU decisions on state aid are structured in clearly separated sections.

- *The Preamble*, containing information on the parties, i.e., on the one hand, the Commission, and on the other hand, a Member state and/or a private party (usually a competitor of the recipient of the stat-aid), the appealed judgment of the

[1] https://github.com/adele-project/demosthenes

Argumentative elements	Tag	Mandatory attributes of the element			Optional attribute of the element		
		Name	Value	Tag	Name	Value	Tag
Premise	<prem>	Identifier	A1, A2, An B1, B2, Bn	ID="An"	/	/	/
		Type	Legal	T="L"	Argumentation scheme	Argument from Rule	S="Rule"
						Argument from Precedent	S="Prec"
						Authoritative Argument	S="Aut"
						Argument from Verbal Classification	S="Class"
						Argument from Interpretation	S="Itpr"
						Argument from Principle	S="Princ"
			Factual	T="F"	/	/	/
Conclusion	<conc>	Identifier	An, Bn, Cn	ID="An"	/	/	/

Table 1: Pre-existing annotation scheme.

Court of First Instance, which is called "General Court", and the composition of the Court;
- *Case background*, including facts and the procedural case history before the General Court;
- *The judgement under appeal*, reporting the assessment of the General Court in the first instance decision;
- *The Appeal*, reporting *The Grounds of Appeal* i.e., reasons why the first instance judgement is challenged. For each ground of appeal, two subsections can be identified: (i) the *Arguments of the Parties*, supporting or attacking the challenge and (ii) the *Findings of the Court*, i.e., the Court's reasoning process, which leads to a decision on the parties' claims;
- *Costs*, i.e., the attribution of costs;
- The *Ruling*, i.e., the final decision and orders to the parties.

In analysing the CJEU decisions, we did not consider sections related to *the preamble*, *the case background*, and *the judgment under appeal*, where no arguments are presented. The same is true with regard to the *costs* and the *final ruling* sections, the latter usually repeating the conclusion of each argument chain and reporting orders to the parties. Since our primary purpose is to capture the argumentative patterns of the CJEU reasoning, we also excluded the section related to the *arguments of the parties*. Thus, the most relevant part is the *Findings of the Court*, reporting all argumentative steps leading to the final ruling. This section is characterised by a set of interacting inferences, which ultimately lead to conclusions on the parties' claims. Each inference links a set of input statements (premises) to an output statement(s) (intermediate or final) conclusion, which may support or attack further inferences.

3.2 The annotation procedure

For the purpose of this study, we added an additional layer of annotation to the Demosthenes dataset, aimed at capturing, for each argument chain, the inferential connection(s) between a set of premises (P_1) and their outcome(s) (P_2). As noted in section 2, a very limited number of works has dealt with predicting argumentative relationships in legal documents.

Compared to the studies by Mochales and Moens [29], Poudyal et al. [37], Teruel et al. [45], Yamada et al. [51], we distinguish between different typologies of support and attack relations. As detailed in the following, we defined 4 main types of links: (i) the *Support fro Premise(s)*; (ii) the *Support from Failure*; (iii) the *Rebuttal*; (iv) the *Undercut*. Additionally, we have established an additional, even though minor, connection, which we called (v) *Rephrase*.

We defined a set of guidelines to annotate the corpus. Both the definition of the guidelines and the labelling process included several revisions. Corrections were also suggested by an analysis of the annotation agreement. The annotation was done at the sentence level by four experts in the legal domain, using periods, semicolons, and line breaks as delimiters. All annotations have been checked, controversial instances have been discussed in a reconciliation phase by two or more expert annotators.

The different connections between a proposition P_1 and the set of propositions P_2 are captured by attaching to P_2 the attribute(s) reported in table 2 and further described below, whose value(s) corresponds to the IDs of the propositions in P_1. Note that multiple propositions can be combined in a single inferential connection, thus playing the same argumentative function. For instance, two premises can jointly support one or more final (or intermediate) conclusion(s), which in turn can be undercut by more than one proposition. Moreover, the identified types of relations are not exclusively between each other. Thus, a single proposition may be assigned more than one type of link, and thus multiple attributes. For instance, a premise may simultaneously challenge the acceptability of an inference between two propositions and support a certain outcome.

To distinguish between the different types of inferential link, we partially relied on (a) recurrent linguistic indicators, including keywords and word patterns; and (b) context indicators, as detailed in the following.

It is important to note that our guidelines were designed to capture the highly complex and heterogeneous argumentative structure of the judicial discourse. However, they are independent of both the text language and the specific legal domain. Thus, they can be easily applied to and tested on different kinds of court decisions on other legal matters.

Support from premise(s). The attribute `SUP` indicates that the propositions in P_2 are the outcome of a set of premises including P_1. Since we focused on the *Finding of the Court* subsection, reporting the Court's reasoning process, this is the most recurrent type of inferential link in the analyzed corpora. It is indeed used by the CJEU to directly justify its decisions on the parties' claims. By analysing the corpora, we identified some recurrent textual indicators signalling

Argumentative elements	Tag	Optional attribute of the element		
		Name	Value	Tag
Premise	<prem>	Type of inferential link	Support from Premise(s)	SUP="An"
			Support from Failure	SFF="An"
			Rebuttal	ATT="An"
			Undercut	INH="An"
			Rephrase	REPH="An"
Conclusion	<conc>	Type of inferential link	Support from Premise(s)	SUP="An"
			Support from Failure	SFF="An"

Table 2: Annotation scheme for inferential links.

the presence of support from the premise(s), which include: *"consequently"*, *"indeed"*, *"accordingly"* *"therefore"*, *"it follows that"*, *"this means that"*, *"in light of the foregoing"*.

As an example, consider the following propositions:

> *<prem ID="D2" T="F" SUP="D3|D4">... However, it must be pointed out that that argument is based on a misreading of that judgment.</prem> <prem ID="D3" T="L" S="Prec|Class">... It is apparent from that judgment that the fact of coming from a compulsory levy is, on the contrary, sufficient to identify State resources (see, to that effect, judgment of 28 March 2019, Germany v Commission, C-405/16 P, EU:C:2019:268, paragraphs 65 to 72).</prem> <prem ID="D3" T="L" S="Prec|Class">... On the other hand, it is irrelevant that the financing mechanism at issue does not, strictly speaking, fall within the category of fiscal levies under national law (see, to that effect, order of 22 October 2014, Elcogás, C-275/13, not published, EU:C:2014:2314, paragraph 31).</prem> <prem ID="D5" T="F" SUP="D2">... The first part of the fourth ground of appeal must therefore be rejected as unfounded.</prem>* (Case C-850/19 P, para 46).

Support from failure. The attribute `SFF`, indicates that the proposition P_2 is (indirectly) entailed by P_1, asserting the failure of the opposing argument P_3 (i.e., a proposition attacking P_2). This inferential connection is used by the Court whenever the burden of proof on P_3 has not been met. Textual indicators signalling support from failure include: "has not shown", "has failed to demonstrate", "it was not such as to prove", etc. As an example, consider the following propositions ():

> *<prem ID="D1" T="F">First, as the Commission correctly contends, the General Court responded in detail to the complaint relating to an alleged breach of the principle of proportionality, raised in the fifth plea in law in the action for annulment, and to that relating to the calculation of the amount of aid to be recovered, raised in the sixth plea in law in that action.</prem> <prem ID="D2" T="F">Secondly, the Hellenic Republic has not indicated with sufficient precision the other complaints put forward by it at first instance to which the General Court did not respond.</prem> <conc ID="D3" SUP="D1" SFF="D2"> In those circumstances, the second part of this ground of appeal must be rejected as, in part, unfounded and, in part, inadmissible.</conc> (Case C-431/14 P, para 80–82).*

Rebuttal. The attribute `ATT` indicates that the proposition P_2 is contradicted by the set of premises P_1. The attacked proposition is a statement by one of the parties, which is mentioned by the Court for the purpose of denying it. In this case, there are no reliable textual indicators. Therefore, the semantics of the concerned propositions has to be carefully examined. As an example, consider the following propositions:

> *<prem ID="A22" T="F" ATT="A24|A25|A27">As regards the first argument of the first part of the first ground of appeal, concerning compensation for a structural disadvantage, Orange relied at first instance on the judgments of the General Court of 16 March 2004, Danske Busvognmænd v Commission (T-157/01, EU:T:2004:76), and of 28 November 2008, Hotel Cipriani and Others v Commission (T-254/00, T-270/00 and T-277/00, EU:T:2008:537), in support of its claim that an advantage eliminating additional burdens which were imposed by derogating arrangements and were not borne by competing undertakings does not constitute State aid.</prem> <prem ID="A23" T="F" ATT="A24|A25|A27">Indeed, according to Orange, compensation for a structural disadvantage may preclude the categorisation of a measure as State aid in certain specific situations, not merely in cases involving services of general public interest.</prem> <prem ID="A24" T="L" S="Class">The General Court rejected that argument in paragraphs 42 and 43 of the judgment under appeal, stating that, even on the assumption that it were established, the compensatory nature of the costs reduction granted in the present case would not make it possible to preclude the categorisation of that measure as State aid.</prem> <prem ID="A25" T="L" S="Prec|Class">The General Court stated in that regard that it is apparent from the case-law of the Court of Justice, in particular paragraphs 90 to 92 of the judgment of 9 June 2011, Comitato Venezia vuole vivere and Others v Commission (C-71/09 P, C-73/09 P and C-76/09 P, EU:C:2011:368), that it is only in so far as a State measure must be regarded as compensation for the services provided by undertakings entrusted with performing a service in the general public interest in order to discharge public service obligations in accordance with the criteria established in the judgment of 24 July 2003, Altmark Trans and Regierungspräsidium Magdeburg (C-280/00, EU:C:2003:415), that such a measure falls outside Article 107(1) TFEU.</prem>... <prem ID="A27" T="F"> Indeed, it is clear that, to date, the only situation recognised by the Court's case-law in which the finding that an economic advantage has been granted does not lead to the measure at issue being categorised as State aid within the meaning of Article 107(1) TFEU is that in which a State measure represents the compensation for the services provided by undertakings entrusted with performing a service in the general public interest in order to discharge public service obligations, in accordance with the criteria established in the judgment of 24 July 2003, Altmark Trans and Regierungspräsidium Magdeburg (C-280/00, EU:C:2003:415).</prem> (Case C-211/15 P, para 40-44)*

Undercut. The attribute `INH`, indicates that the applicability of the proposition P_2 is denied by the set of premises P_1. The undercut proposition is a statement by one of the parties, which is mentioned

by the Court for the purpose of rejecting it. In this case, there are no reliable textual indicators. Therefore, the semantics of the concerned propositions has to be carefully examined. As an example, consider the following propositions:

<prem ID="A3" T="L|F" S="Prec|Itpr" INH="A5|A7|A8|A10"> As regards the arguments alleging that it did not enjoy an economic advantage, Orange submitted before the General Court that it was apparent from the judgment of 23 March 2006, Enirisorse (C-237/04, EU:C:2006:197), that a law which merely prevents an undertaking's budget being burdened with a charge which, in a normal situation, would not have existed, does not confer an advantage on that undertaking for the purpose of Article 107(1) TFEU.</prem> ... <prem ID="A5" T="F">In paragraphs 38 to 41 of the judgment under appeal, the General Court dismissed the argument based on the judgment of 23 March 2006 Enirisorse (C-237/04, EU:C:2006:197), taking the view that that case-law was applicable only in cases involving 'dual derogation' arrangements, that is to say arrangements whereby, in order to prevent the budget of the beneficiary of the measure being burdened with a charge which, in a normal situation, would not have existed, provision is made for a derogation intended to neutralise a previous derogation from the general system in place, which was not the case here.</prem> ... <prem ID="A7" T="L" S="Prec|Class">It should be noted in that regard that, in paragraphs 46 to 48 of the judgment of 23 March 2016, Enirisorse (C-237/04, EU:C:2006:197), the Court held that national legislation which offers an advantage neither to shareholders of a company nor to the company itself, in so far as it merely prevents its budget being burdened with a charge which, in a normal situation, would not have existed and therefore simply regulates an exceptional right and without seeking to reduce a charge which that company would normally have had to bear cannot be regarded as an advantage within the meaning of Article 107(1) TFEU.</prem> <prem ID="A8" T="F" S="Aut">It should be noted, as observed by the Advocate General in point 42 of his Opinion, that a particular feature of the situation which gave rise to the judgment of 23 March 2006, Enirisorse (C-237/04, EU:C:2006:197), was that it concerned a national measure which had the effect of neutralising the effects of a system which derogated from the general system in place.</prem>... <prem ID="A10" T="F" S="A9">It inferred from this that the latter arrangements were not the arrangements normally applicable to France Télécom civil servants, so that the 1996 Law had not removed an abnormal burden borne by the budget of that undertaking or reverted to the normal arrangements.</prem> (Case C–211/15 P, para 23–29).

Rephrase. The attribute REPH, indicates that P_1 and P_2 rephrase each other, i.e., they express the same content. As an example, consider the following propositions:

<prem ID="C2" T="F" REPH="C7">It is apparent from paragraphs 124 to 130 and from paragraph 200 of the judgment under appeal that the appellants submitted that the tax scheme established in Articles 17 and 18 of Law 342/2000 was extended only in respect of companies and entities which took part in transfers of assets in exchange for shares under the system of fiscal neutrality provided for in Article 7(2) of Law 218/1990.</prem> <prem ID="C3" T="F" REPH="C6">They argued that, following the entry into force of Legislative Decree 344/2003 and the establishment of the special exemption regime known as 'shareholding exemption', the risk of double taxation disappeared for the companies and entities which carried out such transactions under the system of fiscal neutrality provided for in Article 4 of Legislative Decree 358/1997.</prem> ... <prem ID="C6" T="F" REPH="C3">In that context, the applicants submit essentially ... that the 2003 tax reform eliminated any risk of the double taxation of the gains arising on the transfer of assets under the system of fiscal neutrality established by Article 4 of Legislative Decree ... 358/1997, namely the taxation of both the companies transferring and the companies in receipt of assets.</prem> <prem ID="C7" T="F" REPH="C2">On the other hand, that reform did not remove the risk of the double taxation of the gains arising in connection with the transfer of assets under the system of fiscal neutrality introduced by Law ... 218/1990. The [appellants] maintain that that explains the decision of the Italian legislature to extend the realignment scheme under Articles 17 and 18 of Law ... 342/2000 only to assets transferred in the context of Law ... 218/1990.</prem> (Case C–452/10 P, para 97 and 100).

3.3 Inter-Annotator Agreement

The agreement was measured on 14 documents tagged by 2 annotators. To calculate the agreement on the presence of the links, we considered each pair of sentences labelled as premises or conclusions and treated the problem as a binary classification. We obtained a Cohen's κ [10] of 0.59, which indicates a good agreement. More details are reported in Table 3.

Following Teruel et al. [45], we also computed the agreement on the link type by only considering pairs where the two annotators agreed on the presence of a link, reaching the score of κ=0.57. This result is mainly due to the disagreement regarding the classes SFF, INH, and REPH. In particular, all SFF and REPH in which there is disagreement have been labeled SUP by the other annotator, while the only INH has been labeled ATT. These classes, along with the ATT class, are strongly underrepresented in the document, significantly increasing the difficulty of the task. Indeed, despite the low κ, the two annotators agreed in 95% of the cases. Moreover, these tasks require extensive knowledge of the domain and an interpretation of the text in order to establish a connection with the argumentation model followed by the judges who authored the documents. Therefore the relatively low agreement, which is a limitation of this part of our dataset, could be partly explained as an effect of the complexity of the task at hand, as well as of the scarceness of the three mentioned classes. To mitigate the low Cohen's K possible solutions would be (a) to retrain taggers for specific complex issues, and (b) to merge the SFF with SUP, since the former can be subsumed by the broader support category. On the other hand, we shall remark that low agreement in link prediction and relation classification is a common issue affecting several works in argumentation mining in the law domain [45] (see also Section 2). For example, a similar agreement score was obtained by Kirschner et al. [22], but they considered only links between nearby elements.

3.4 Corpus statisics

The final composition of the dataset is reported in Table 4. The average number of links per document is 65, with a maximum of 152 and a minimum of 22. Most links (33%) connect adjacent argumentative components, while 73% links connect components with a distance in the range [-5,5]. Expanding the range to [-10, 10], increases the number of links to the 88%.

	Link	Type				
		SUP	SFF	ATT	INH	REPH
Only Ann. 1	241	5	10	0	1	1
Only Ann. 2	203	11	2	1	0	3
Both Ann.	325	296	8	4	0	0
κ	0.59			0.57		

Table 3: Number of links and link types tagged by each annotator and agreement between them.

Element	No.
documents	40
sentences	9320
prem	2375
conc	160
links	2595

Relation	No.	% of total links
SUP	2257	86.98
SFF	93	3.58
ATT	143	5.51
INH	36	1.39
REPH	66	2.54

Table 4: Dataset statistics.

It is evident that the distribution of relationship classes is strongly unbalanced. Support relations represent more than 85% of the links, while other types represent less than 6% each.

4 EXPERIMENTAL SETTING

Following previous work [12, 13, 27], we address the link prediction task as a binary classification problem over textual inputs. Given two inputs, the *source* and the *target*, that belong to the same document and are known to be argumentative, the task's objective is to predict whether there is a link from the source to the target. We refer to source-target input pairs as *argumentative pairs*.

4.1 Data Preparation

The length of the documents and the high number of argumentative elements inside of them inherently lead to a large number of possible argumentative pairs, especially the negative ones, as their number grows exponentially. For these reasons, considering every argumentative pair is computationally demanding for training a classification model. Moreover, related work on argumentative link prediction [12, 22, 37] observes that the majority of argumentative pairs are between inputs that are relatively close to each other. Thus, many studies only consider argumentative pairs whose input distance is below an arbitrary threshold to significantly reduce the number of possible pairs.

Following Galassi et al. [12], we define spatial distance as the number of argumentative sentences between the source and the target, using a positive value when the source precedes the target and a negative otherwise. We experiment with two distinct training settings according to the chosen distance threshold: W_{small} and W_{large}. We allow argumentative pairs with input distance in the range $[-3, 7]$ and $[-6, 14]$ in W_{small} and W_{large}, respectively. These ranges have been arbitrarily established based on the statistics observed in the training set. The former interval includes 77% of the links in the training sets, while the latter 90%. Each of these settings defines a subset of our extended dataset, whose details are reported in Table 5.

Applying an arbitrary distance threshold has the drawback of excluding long-distance argumentative pairs. However, their limited number should not affect model performance significantly. To better evaluate the impact of this choice, we test the models both on the original test set, which includes all possible pairs, and the test sets where the distance constraint is applied.

We split our extended version of the Demosthenes dataset into training, validation, and test sets for model evaluation. In particular, we split textual inputs such that all inputs belonging to the same document are in the same split and each document is randomly placed in one of the splits. Table 5 reports dataset statistics concerning built splits.

4.2 Models

We consider two neural classifiers, transformer-based and residual attention networks [14], that are widely adopted for argumentative link prediction. Moreover, we consider a uniform random classifier and a majority classifier as standard baselines.

4.2.1 ResAttArg Ensemble. We use the architecture proposed by Galassi et al. [14], which we refer to as ResAttArg. The model comprises stacks of dense and Long Short-Term Memory (LSTM) [20] layers, a co-attention module [13], and residual connections [19]. Moreover, it uses 300-dimensional GloVe[34] pre-trained embeddings for text representation. Regarding distance, the model is designed to encode the distance between the source and the target as a 10-bit array. Figure 1 illustrates the network architecture. We maintain the same hyper-parameters of the original paper, resulting in a network with about 100,000 trainable parameters.

The model is designed to be jointly trained over three different tasks: link prediction, relation classification, and component classification. In particular, given an argumentative pair, the model outputs: (i) the argumentative type of the source and target, (ii) whether there is a link from source to target, and (iii) the link relation type.

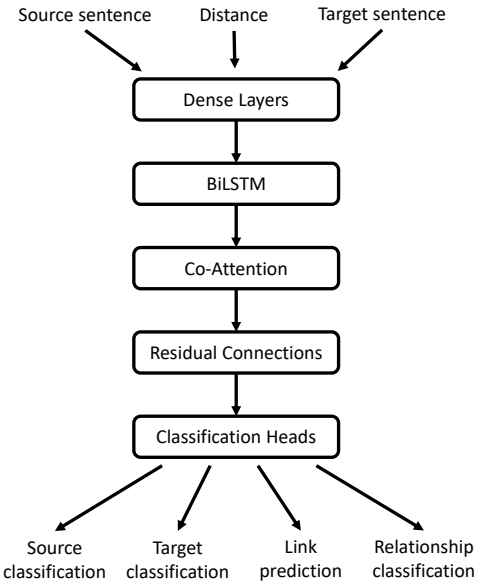
Following [14], we repeat the training multiple times training ten models. In the evaluation, we consider both the average performance of every single model and the result obtained by their ensemble. The ensemble thus created can be considered as a model with 1 million parameters.

4.2.2 DistilRoBERTa. For the transformer-based classifier class, we consider DistilRoBERTa [39], a distilled version of RoBERTa [25]. This distillation process allows reducing the model size up to 40% while retaining 97% of its language understanding capabilities and being 60% faster. Similarly to BERT [11], RoBERTa is trained with the masked language modelling objective for language understanding: given a sentence, a random word mask with 15% masking probability is applied. Then, the model is trained to predict the original word for each masked one. This pre-training process allows the model to learn the semantic and syntactic properties of the language. Subsequently, the model can be fine-tuned to address specific tasks in the same language. Differently from BERT, RoBERTa is pre-trained on a larger dataset of English texts, it is trained longer and on longer sequences, it is calibrated with a more extensive hyper-parameter search, and removes the next sentence prediction training objective.

	Split	Doc.	*prem*	*conc*	Original *link*	Original *no-link*	W_{small} *link*	W_{small} *no-link*	W_{large} *link*	W_{large} *no-link*
(a)	Train	20	1096	66	1169	77491	901	10029	1049	19651
	Validation	10	500	37	539	31791	429	4601	511	8969
	Test	10	779	57	887	80163	679	7341	785	14675
	Total	40	2375	160	2595	189445	2009	21971	2345	43295
(b)	Train w/ oversampling	20	1096	66	-	-	9911	10029	18882	19671

Table 5: (a) Split sets composition in the Original, W_{small} and W_{large} training settings. (b) Train set composition with positive class oversampling.

Figure 1: ResAttArg architecture.



We train RoBERTa on the task of link prediction, providing the two sentences as input. We experiment with the DISTILROBERTA-BASE Huggingface[2] model version, which has 82 million parameters.

4.3 Oversampling

Even when applying a distance threshold to reduce the number of argumentative pairs, the ratio of positive and negative pairs is significantly in favour of negative ones. Such a class imbalance may negatively affect model training. We study the impact of class imbalance and propose positive class oversampling to mitigate its effect. More precisely, we oversample positive argumentative pairs in the training set by a 9 and 17 factor in the W_{small} and W_{large} training settings, respectively. The last row of Table 5 reports the composition of the oversampled training sets.

[2]https://huggingface.co/

5 RESULTS AND DISCUSSION

We report the results concerning the link prediction Table 6. The best result on the link prediction task over the entire test set is obtained by the ensemble of residual attentive networks trained with the W_{large} setting. It reaches an F1 score of 0.41 on the positive class and a macro average of 0.70. This result shows a huge improvement with respect to the random baseline, which scores 0.02 on the positive class and 0.34 on the macro. When trained in the W_{small} setting, the model reaches almost identical results. Considering the average performance of the single networks leads to worse performances, obtaining about 6 and 9 percentage points less than the ensemble trained on the W_{large} and W_{small} settings, respectively. It is interesting to notice that oversampling significantly worsens the performance, probably due to the overfitting on the training set.

DistilRoBERTa obtains results comparable to the majority baseline, predicting the negative class in almost all cases. The use of oversampling improves the model, but the result remains unsatisfactory, reaching an F1 score of 0.08 and 0.09 on the positive class in the W_{small} and W_{large} settings respectively.

With respect to the subsets defined by the W_{small} and W_{large} setting, we can see that DistilRoBERTa greatly improves, reaching gaining more than 0.20 points in the W_{small} setting and 0.10 in the W_{large} one. The residual models improve as well, although it gains about half the point of DistilRoBERTa. Moreover, we can observe that oversampling does not affect the final result in the W_{small} setting, while it slightly improves in the W_{large} one.

These observations suggest that the remarkable difference between the performance of the two models may be due to the superior ability of the residual network to generalize and be robust against the unbalance in the training set, especially when used in ensemble.

Since the residual model jointly performs also component classification, we report its results in Table 7. The residual model is slightly better than the results presented by Grundler et al. [16] using random forests over TF-IDF encodings, improving the macro F1 score by 1 percentage point. It is important to remark that the two results are not directly comparable since they were obtained with two different evaluation methods: Grundler et al. [16] use a cross-validation setting instead of splitting the dataset into train and test splits. We also experiment with the relation classification task, but the models do not yield satisfactory results. They always predict the majority class (SUP), performing similarly to the majority baseline.

Training setting	Model	Test setting: Original			Test setting: W_{small}			Test setting: W_{large}		
		link	*no-link*	Avg.	*link*	*no-link*	Avg.	*link*	*no-link*	Avg.
	Majority baseline	0.00	0.99	0.50	0.00	0.96	0.48	0.00	0.97	0.49
	Random uniform baseline	0.02	0.66	0.34	0.14	0.65	0.40	0.09	0.66	0.37
W_{small}	DistilRoBERTa	0.00	0.99	0.50	0.00	0.96	0.48	-	-	-
	w/ oversampling	0.08	0.95	0.51	0.32	0.92	0.62	-	-	-
	ResAttArg	0.31	0.99	0.65	0.44	0.95	0.69	-	-	-
	w/ oversampling	0.23	0.98	0.60	0.44	0.94	0.69	-	-	-
	ResAttArg (Ensemble)	0.40	0.99	0.69	**0.49**	0.96	0.73	-	-	-
	w/ oversampling	0.29	0.99	0.64	**0.49**	0.95	0.72	-	-	-
W_{large}	DistilRoBERTa	0.00	0.99	0.50	-	-	-	0.00	0.97	0.49
	w/ oversampling	0.09	0.97	0.53	-	-	-	0.20	0.95	0.57
	ResAttArg	0.35	0.99	0.67	-	-	-	0.40	0.97	0.68
	w/ oversampling	0.25	0.99	0.62	-	-	-	0.39	0.96	0.67
	ResAttArg (Ensemble)	**0.41**	0.99	0.70	-	-	-	0.45	0.98	0.71
	w/ oversampling	0.34	0.99	0.67	-	-	-	**0.47**	0.97	0.72

Table 6: Results for the link prediction task. We report the F1 score for the positive and negative class, along with their macro average. The rows represent the trained models, grouped by the training setting. Columns represent the test set in the three different settings.

Training setting	Model	*prem*	*conc*	Avg.
	Grundler et al. [16]	0.99	0.77	0.88
W_{small}	ResAttArg	0.98	0.77	0.87
	w/ oversampling	0.98	0.80	0.89
W_{large}	ResAttArg	0.98	0.80	0.89
	w/ oversampling	0.99	0.80	0.89

Table 7: Results for the component classification task. We report the F1 score for each class along with their macro average.

6 CONCLUSION

Argumentative link prediction is a crucial task in the legal domain, which has only been partially addressed so far. We contributed to it with a study on the CJEU on fiscal state aid. In particular, we designed a novel annotation scheme for legal arguments in these decisions. We focus on the relations between propositions, i.e., argument components, to capture, for each argument chain, the inferential connection(s) between a set of premises and their outcome(s). We used such a scheme to add an annotation layer to the *Demosthenes* corpus by Grundler et al. [16], thereby building a new dataset. Compared to previous works in this area, we distinguished between different typologies of support and attack relations, i.e., (i) Support form Premise(s), (ii) Support from Failure, (iii) Rebuttal, and (iv) Undercut. We also established the additional (v) Rephrase relation.

Our new dataset enabled us to run an empirical study, where we compared two architectures (DistilRoBERTa and Ensemble ResAttArg) and studied their performance by changing a number of hyperparameters (including, importantly, the distance between arguments when evaluating links) and training regimes (with/without oversampling). The best results were obtained using ResAttArg without oversampling. The distance between arguments does not play a crucial role in training. This is very interesting since it shows the robustness of the approach and the practical viability of our proposed solution as residual networks are less resource-intensive than transformer-based architectures. In fact, there are almost two orders of magnitude between ResAttArg and DistilRoBERTa.

Our study also confirms that link prediction is one of the most challenging AM tasks, hampered by data scarcity and low annotator agreement, confirming the complexity of the task itself. Reconstructing the highly complex argumentative structures of Courts' decisions in the wild, i.e., in natural language texts, is error-prone because of the existing gap between the judicial discourse as expressed in natural language and the typical argumentation patterns analyzed in theoretical studies. How to fill this gap remains an open research challenge, requiring the effort of the AM community.

An additional level of difficulty is the distance that may be between linked arguments. This makes recognizing links challenging not only for human annotators, who need to work with a larger document span, but also for NLP systems, as the number of possible pairs that may or may not be linked grows quadratically with the maximum distance among them, and so does the ratio between negative and positive examples. This is confirmed by our empirical results, which indicate that increasing the scope of the analysis (maximum distance between linked arguments) does not cause results to improve. This is one aspect that deserves further attention.

Other directions of future work include expanding our dataset and applying our annotation scheme to judicial decisions on different legal matters. We also want to study the impact of different degrees of oversampling, undersampling, and data augmentation [35]. Finally, we plan to explore relation classification by addressing the problem of under-representation of classes other than support through data augmentation or by merging them into a single class.

ACKNOWLEDGMENTS

This work has been partially funded by EU H2020 under grant agreement 101017142 (StairwAI), by EU Justice under grant agreement 101007420 (ADELE), by EU H2020 ERC under grant agreement 833647 (CompuLaw), by Italian Ministry of Education and Research's PRIN programme under grant agreement 2017NCPZ22 (LAILA), by European Commission NextGeneration EU programme, PNRR - M4C2 - Investimento 1.3, PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI".

REFERENCES

[1] Kevin D. Ashley, Ravi Desai, and John M. Levine. 2002. Teaching Case-Based Argumentation Concepts Using Dialectic Arguments vs. Didactic Explanations. In *Intelligent Tutoring Systems*. 585–595.

[2] Katie Atkinson and Trevor J. M. Bench-Capon. 2019. Reasoning with Legal Cases: Analogy or Rule Application?. In *ICAIL*. ACM, 12–21.

[3] Katie Atkinson and Trevor J. M. Bench-Capon. 2021. Argumentation schemes in AI and Law. *Argument Comput.* 12, 3 (2021), 417–434.

[4] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *EACL*. 251–261.

[5] Trevor J. M. Bench-Capon, James B. Freeman, Hanns Hohmann, and Henry Prakken. 2004. Computational Models, Argumentation Theories and Legal Practice. In *Argumentation Machines*. Vol. 9. Springer, 85–120.

[6] Trevor J. M. Bench-Capon, Henry Prakken, and Giovanni Sartor. 2009. Argumentation in Legal Reasoning. In *Argumentation in Artificial Intelligence*. Springer, 363–382.

[7] Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *IJCAI*. 5427–5433.

[8] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument Pair Extraction from Peer Review and Rebuttal via Multi-task Learning. In *EMNLP*. ACL, Online, 7000–7011.

[9] Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *EMNLP*. ACL, 1374–1379.

[10] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20 (1960), 37 – 46.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. ACL, 4171–4186.

[12] Andrea Galassi, Marco Lippi, and Paolo Torroni. 2018. Argumentative Link Prediction using Residual Networks and Multi-Objective Learning. In *ArgMining@EMNLP*. Association for Computational Linguistics, 1–10.

[13] Andrea Galassi, Marco Lippi, and Paolo Torroni. 2021. Attention in Natural Language Processing. *IEEE TNNLS* 32, 10 (2021), 4291–4308. https://doi.org/10.1109/TNNLS.2020.3019893

[14] Andrea Galassi, Marco Lippi, and Paolo Torroni. 2021. Multi-Task Attentive Residual Networks for Argument Mining. *CoRR* abs/2102.12227 (2021).

[15] Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. 2015. Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. In *ICAIL*. 69–78.

[16] 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 *ArgMining@COLING*. 143–157.

[17] Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Christoph Burchard, et al. 2022. Mining legal arguments in court decisions. *arXiv preprint arXiv:2208.06178* (2022).

[18] Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law* 14 (2006), 305–345.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80.

[21] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. 2009. LKIF Core: Principled Ontology Development for the Legal Domain. In *Law, Ontologies and the Semantic Web*, Vol. 188. IOS Press, 21–52.

[22] Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications. In *ArgMining*. ACL, Denver, CO, 1–11.

[23] John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45, 4 (2020), 765–818.

[24] Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. *ACM Trans. Internet Techn.* 16, 2 (2016), 10:1–10:25.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019).

[26] Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019. The Evolution of Argumentation Mining: From Models to Social Media and Emerging Tools. *Inf. Process. Manage.* 56, 6 (nov 2019), 22 pages.

[27] Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2021. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artif. Intell. Medicine* 118 (2021), 102098.

[28] Raquel Mochales and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In *JURIX*. 11–20.

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

[30] 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 (2007), 89.

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

[32] Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument Mining with Structured SVMs and RNNs. In *ACL*. ACL, Vancouver, Canada, 985–995.

[33] Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *ICAIL*. 98–107.

[34] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*. Doha, Qatar, 1532–1543.

[35] Sezen Perçin, Andrea Galassi, Francesca Lagioia, Federico Ruggeri, Piera Santin, Giovanni Sartor, and Paolo Torroni. 2022. Combining WordNet and Word Embeddings in Data Augmentation for Legal Texts. In *NLLP*. 47–52.

[36] Prakash Poudyal, Teresa Gonçalves, and Paulo Quaresma. 2019. Using Clustering Techniques to Identify Arguments in Legal Documents. In *ASAIL@ICAIL*.

[37] Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: legal corpus for argument mining. In *ArgMining*. 67–75.

[38] Henry Prakken and Giovanni Sartor. 1996. A Dialectical Model of Assessing Conflicting Arguments in Legal Reasoning. *Artif. Intell. Law* 4, 3-4 (1996), 331–368.

[39] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[40] Muhammad Tawsif Sazid and Robert E. Mercer. 2022. A Unified Representation and a Decoupled Deep Learning Architecture for Argumentation Mining of Students' Persuasive Essays. In *ArgMining*. 74–83.

[41] Lida Shi, Fausto Giunchiglia, Rui Song, Daqian Shi, Tongtong Liu, Xiaolei Diao, and Hao Xu. 2022. A Simple Contrastive Learning Framework for Interactive Argument Pair Identification via Argument-Context Extraction. In *EMNLP*. ACL, Abu Dhabi, United Arab Emirates, 10027–10039.

[42] Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*. 46–56.

[43] Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43, 3 (Sept. 2017), 619–659.

[44] Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies* 11, 2 (2018), 1–191.

[45] Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Legal text processing within the MIREL project. In *Workshop on Language Resources and Technologies for the Legal Knowledge Graph*. 42.

[46] Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines. In *LREC*.

[47] Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

[48] Joel P Trachtman. 2013. The tools of argument: How the best lawyers think, argue, and win. *Argue, and Win (July 29, 2013)* (2013).

[49] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

[50] Huihui Xu, Jaromír Šavelka, and Kevin D Ashley. 2020. Using Argument Mining for Legal Text Summarization. In *JURIX*. 184–193.

[51] Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2019. Neural network based rhetorical status classification for japanese judgment documents. In *Legal Knowledge and Information Systems*. IOS Press, 133–142.

[52] Jian Yuan, Zhongyu Wei, Yixu Gao, Wei Chen, Yun Song, Donghua Zhao, Jinglei Ma, Zhen Hu, Shaokun Zou, Donghai Li, and Xuanjing Huang. 2021. Overview of SMP-CAIL2020-Argmine: The Interactive Argument-Pair Extraction in Judgement Document Challenge. *Data Intelligence* 3, 2 (06 2021), 287–307.

[53] Gechuan Zhang, Paul Nulty, and David Lillis. 2022. Enhancing Legal Argument Mining with Domain Pre-training and Neural Networks. *CoRR* abs/2202.13457 (2022).