Legal Argument Mining: Recent Trends and Open Challenges

Rūta Liepiņa[1,*], Francesca Galloni[1], Francesca Lagioia[1,2], Marco Lippi[3], Mariaceleste Musicco[1], Burcu Sayin[4], Andrea Passerini[4] and Giovanni Sartor[1,2]

[1]ALMA-AI, CIRSFID, Department of Law, University of Bologna, Bologna, Italy
[2]Law Department, European University Institute, Florence, Italy
[3]Department of Information Engineering (DINFO), University of Florence, Florence, Italy
[4]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Abstract

This paper presents a brief survey of recent trends in legal argument mining, focusing on the early use of large language models in this subfield. As legal texts, especially judicial decisions, increase in volume and complexity, the need for effective tools to extract and analyse legal arguments becomes more pressing. The paper outlines key datasets and tasks in legal argument mining, and identifies challenges and open issues to guide future research.

Keywords

Legal argument mining, large language models, argumentation datasets, survey

1. Introduction

The growing complexity and volume of legal texts – particularly judicial decisions – has spurred interest in computational tools capable of capturing and organising legal reasoning [1]. Argument mining (AM) has emerged as a crucial task for the legal domain, to extract and analyse argumentative structures from textual documents [2][3]. Indeed, AM can be highly beneficial to both legal research and legal decision making, improve judicial transparency and contribute to the advancement of AI-assisted reasoning. Judicial decisions are inherently argumentative, often justifying legal conclusions through layered reasoning, with references to precedents and legislation. Structured access to this information can empower legal professionals and scholars to better understand, critique, and apply legal rulings.

The information acquisition bottleneck has significantly affected progress in legal argument mining (LAM). Traditional approaches to information extraction have faced persistent challenges in addressing LAM due to aspects such as the complexity of legal language, its context dependency, the need for domain expertise, and variation across jurisdictions [4][5].

The emerging capabilities of large language models (LLMs) [6] hold promise for assisting with LAM. Their capacity for zero-shot and few-shot learning and for contextual understanding can be deployed in identification and or classification of legal content. Their proficiency in reasoning tasks makes them especially promising for AM in law, where subtle distinctions and interpretive nuances have to be taken into account to detect and distinguish patterns of reasoning.

This study aims to map the state-of-the-art research on the use of LLMs for argument mining in judicial decisions and to identify research gaps and open research questions. By focusing on LAM and LLMs, our work aims to serve as a reference for researchers and legal professionals seeking to understand and advance the use of AI in the analysis of judicial reasoning. The paper is structured as

Proceedings of the First Argument Mining and Empirical Legal Research Workshop (AMELR 2025), June 20, 2025, Chicago, United States

*Corresponding author.
ruta.liepina@unibo.it (R. Liepiņa); francesca.galloni5@unibo.it (F. Galloni); francesca.lagioia@unibo.it (F. Lagioia); marco.lippi@unifi.it (M. Lippi); mariaceleste.musicco@unibo.it (M. Musicco); burcu.sayin@unitn.it (B. Sayin); andrea.passerini@unitn.it (A. Passerini); giovanni.sartor@unibo.it (G. Sartor)
0000-0002-2417-3219 (R. Liepiņa); 0009-0001-3727-6356 (F. Galloni); 0000-0001-7083-3487 (F. Lagioia); 0000-0002-9663-1071 (M. Lippi); 0009-0009-5815-4313 (M. Musicco); 0000-0001-6804-127X (B. Sayin); 0000-0002-2765-5395 (A. Passerini)

follows. Section 2 describes the criteria for selecting the relevant studies. Sections 3 and 4 respectively presents an overview of the employed datasets and the argument mining tasks. Section 5 discusses challenges, open questions and outlines future research directions.

2. Mapping the latest advancements: selection criteria

We conducted a targeted literature review to map recent advancements in the use of classical transformer models (e.g., LegalBERT) and LLMs for LAM. The search covers leading databases and venues in AI, computational linguistics, and legal informatics, including the ACM Digital Library, ACL Anthology, JURIX, ICAIL, the Argument Mining Workshop, and the Artificial Intelligence and Law Journal. We used focused keywords – "argument mining", "legal argument mining", "large language models", and "legal argumentation" – to identify studies at the intersection of advanced NLP and legal reasoning. Papers are selected for their methodological, experimental, and dataset-focused relevance to judicial contexts. Three inclusion criteria are applied. *Topical Relevance*, requires significance to the methodological, experimental, and dataset-oriented aspects. *Venue Quality* restricts the selection to reputable, peer-reviewed conferences, journals, or workshops within AI and law or closely related fields. *Temporal Relevance* considers works published from 2020 onward, a period marked by the rapid evolution and widespread adoption of most advanced transformer models and LLMs. Exclusion criteria filter out studies with limited pertinence to LAM, those lacking data analysis, and grey literature (e.g., non-peer-reviewed or self-published works). The final selection includes 21 papers (7 of which use LLMs, and 14 use classical transformer models). We collected citation metadata and examined datasets, distinguishing between new corpora and reused datasets, as well as jurisdictions, domains, and languages. We categorised the AM tasks, and documented the employed models. Contributions are assessed in terms of classical metrics and expert-based legal evaluations. Finally, we highlight recurring and new challenges, which include data limitations, annotation subjectivity, and cross-jurisdictional generalisability.

3. Datasets

Despite growing interest in computational legal reasoning, open-access and high-quality datasets for LAM remain limited, especially when compared to general-domain or even other specialised NLP areas. Nonetheless, several recent contributions have introduced or repurposed datasets for tasks involving legal argumentation, enabling experimentation with classical transformer-based models and LLMs. This section reviews the most relevant corpora in the selected literature, outlining jurisdictions, legal domain focus, language and scope. A comparative table summarising dataset properties and usage is given in Appendix A.

European courts. The supranational nature of European courts, particularly the European Court of Human Rights (ECHR) and the Court of Justice of the European Union (CJEU), has motivated the creation of annotated datasets for LAM. The ECHR in particular has been a major focus. One of the earliest contributions comes from Poudyal et al. [7], who created a corpus of 42 ECHR judgments in English, annotated at the clause level as premise, conclusion, or non-argumentative, along with mapped relations between premises and conclusions. Clauses may serve multiple argumentative roles, reflecting the layered nature of legal reasoning. Expanding this foundation, the LAM:ECHR corpus [8] offers a fine-grained annotations over 373 ECHR judgments in English, focusing on Articles 3, 7, and 8 of the EU Charter of Fundamental Rights. Argument spans are labelled using a Toulmin-inspired model, capturing both argument types (e.g., different methods of interpretation, tests of the principle of proportionality, precedents, and others) and actors (e.g., the ECHR itself, the State, applicants).

Further extending ECHR coverage, Chlapanis et al. [9] exploit the legal reasoning abilities of LLMs. Given a set of arguments, the goal is to predict the next correct statement. To this end, they produced LAR-ECHR, by combining three pre-existing corpora [10][8][11]. The final corpus contains 403 samples from 191 cases in English, each sample including a target argument (i.e., the correct next legal statement)

and several distractors (i.e., incorrect next statements). All targets are extracted from legal reasoning authored by judges, and pertain to the court's application of law to facts and follow parties' submissions. Distractors are selected from the same corpus to match target arguments in style and vocabulary, avoiding paraphrases. The LaCour! corpus [12] introduces oral legal discourse in English and French. It includes 154 transcribed ECHR courtroom dialogues held between 2012 and 2021. Transcripts are annotated with sentence-level labels, identifying questions, opinions (e.g., dissenting, concurring), speaker roles, language and timestamps. Each hearing is linked to the corresponding final judgment, offering a multimodal perspective on case deliberation.

Finally, Demosthenes[13] includes 40 English judgments by the Court of Justice of the European Union (CJEU) on fiscal State Aid, ranging from 2000 to 2018. It focuses on the "Findings of the Court" section. The annotation follows a three-level hierarchical scheme: (1) argumentative components (premises, conclusions), (2) premise types (legal or factual), and (3) argumentation schemes (e.g., Rule, Precedent, Authority). An extension to the dataset[14] introduced inferential relations between components, such as support, rebuttal, and undercut, allowing for rich modelling of argumentative structures.

National courts - common law. Academic researchers have created several datasets that highlight the evidential reasoning in common law jurisdictions. In the U.S., the BVA dataset [15] contains 30 Veterans' Appeals decisions on PTSD claims in English, from 2013 to 2016. The corpus is annotated with three rhetorical roles, relevant to evidentiary reasoning: evidence, reasoning, and findings - mapping directly to how tribunals evaluate and adjudicate claims.

In Canada, a legal argumentation corpus was created from over 28,733 case-summary pairs sourced from CanLII [16]. Initial annotation involved 574 randomly selected summaries [17], later expanded to 1,049. Each case is annotated with the components of an argument triple: issue, conclusion, and reason. The corpus contains statements extracted from full-text decisions and their corresponding summaries. The total number of statements from full texts is significantly higher compared to those extracted from summaries

One of the largest-scale resources to date is the Indian Supreme Court corpus by Ali et al. [18] covering 30,034 English decisions from 1952 to 2012. Through rule-based methods, relational sentence pairs are automatically identified and labelled as Support or Attack, enabling large-scale analysis of argumentative dynamics. The authors later expand the dataset [19], specifically focusing on industrial disputes. Each argument is represented by its start and end sentence, where each sentence in between is considered as a part of the argument. For each argument, annotators were also required to identify the sentence containing the major claim.

Building on previous works [20], Bambroo et al.[21] created two additional English corpora focused on rhetorical role classification: the DIN dataset, based on 150 Indian Supreme Court decisions (including 100 newly annotated cases), and the DUK dataset, covering 50 judgments by the UK Supreme Court. Each sentence is labelled using a seven-role schema – Facts, Lower Court Ruling, Argument, Ratio, Statute, Precedent, Present Court Ruling – reflecting the internal structure of legal decisions.

National courts - civil law. In civil law jurisdictions, legal judgments often follow a codified structure, and argumentation is more tightly bound to statutory interpretation. New datasets from Germany, Italy, and Spain mark a turning point in making these systems accessible for empirical analysis.

The German dataset [22] focuses on proportionality arguments in constitutional law. It includes 300 randomly selected decisions by the German Federal Constitutional Court (GFCC), issued between 1951 and 2021. Annotations, limited to the "merit" section, are based on the GFCC's four-step proportionality test – legitimate aim, suitability, necessity, balancing – and allow multiple labels per sentence.

The Italian dataset [23] consists of 225 VAT rulings in Italian by Regional Tax Commissions from 2010 to 2022, sourced from the Giustizia Tributaria database. Annotations follow a three-level hierarchy – i.e., argument components, premise type, and argument scheme – defined by [13] for the Demosthenes corpus.

The Spanish dataset [24], focuses on family law issues (e.g., child custody, alimony, house allocation).

It includes 3,047 decisions from provincial and higher courts, issued between 2015 and 2020, and sourced from the Spanish Centre for Judicial Documentation (CENDOJ). Annotations cover the following elements: request types (e.g., custody, alimony), judicial principles, factual or legal justifications and decisions. To optimise training, annotated segments were selected based on two criteria: their clarity in representing the target category and their self-contained interpretability.

4. Argument Mining Tasks

LAM encompasses a constellation of computational tasks aimed at dissecting and reconstructing the argumentative elements of judicial decisions. At its core, argument mining seeks to render the complex reasoning processes embedded in legal texts into structured, machine-interpretable representations. This is typically approached through a multi-stage pipeline, aimed at extracting natural language arguments and their relations from textual documents [25][3]. Each stage addresses a sub-task of the problem. At first, usually argumentative sentences (i.e. those containing an argument or part thereof) are detected. Then the boundaries of the various argument components are identified and their characteristics are specified (e.g, distinguishing between premises and conclusions, argument schemes, actors) [26][5][27]. Finally, relationships are predicted between these components and/or between the arguments they are part of [2][3]. We structure our review, distinguishing these tasks into four general categories: argument component detection and classification, structure and relation modelling, legal reasoning, and multi-task and hybrid setups. Additional details on the tasks and employed models can be found in Appendix B.

Detection and classification: At the foundational level of AM pipelines lies the task of identifying argumentative content – distinguishing those statements that contain claims, premises or other reasoning elements from non-argumentative texts. Once identified, these components are further categorised according to their argumentative function (e.g., premise, conclusion) and attributes (e.g., factual vs. legal, actor attribution and rhetorical role). To extract non-overlapping, contiguous sentence spans representing complete legal arguments, Ali et al. [19] identify a set of argument markers, including claim sentences and their supporting premises. They model the task as a text segmentation problem over entire court judgments, integrating local classifications using Integer Linear Programming to produce an optimal document-level segmentation of legal arguments.

Al Zubaer and colleagues [28] focus on the binary classification of argumentative components as premise(s) or a conclusion(s). While, Grundler et al. [13] exemplify the AM layered pipeline through a 4 step framework: (i) detection, i.e., classifying sentences as argumentative or not; (ii) classification as premise(s) or conclusion(s); (iii) type classification of premises as legal, factual or both; and (iv) scheme classification. The last two are multi-label tasks assigning types and argument schemes to legal premises. Muñoz-Soro and others [24] identify relevant argumentative components in family law through binary classification (argumentative/non-argumentative) – with a particular focus on child custody cases – and then introduce a tailored multi-label classification of such sentences, categorised as types of plaintiff's requestions, legal justifications (the main arguments used by the court in custody proceedings), and the court's decisions.

Rhetorical roles are a relevant point of focus too. Walker et al [15], for instance, distinguish among (i) evidence (e.g., describing medical records and lay testimony); (ii) evidence reasoning (explaining how the tribunal interprets such evidences); and (iii) findings (stating formal factual conclusions reached by the decision-maker). Several studies have focused on the Issue–Reason–Conclusion (IRC) framework, to model argumentative structure in legal texts. Xu et al [17] rely on the IRC taxonomy to classify statements according to their argumentative roles. They [16] refine the IRC approach, by incorporating a token-level and subsequent sentence-level classification of (i) court-issues; (ii) conclusions on such issues (court's decision); and (iii) the court's reasons for so concluding, coupling the binary classification task with abstractive summarisation (all unannotated sentence are treated as non-IRC sentences). Similarly, with the ultimate goal of improving the summarisation task, Elaraby and Litman [29] adopt a

multi-class approach for a sentence-level classification according to one of three legal argument roles under the IRC framework. An additional class is assigned to non-argumentative sentences.

Bambroo et al [21] provide a multi-class classification of statements into seven predefined rhetorical roles (Facts, Ratio, Precedent, etc.). Lüders and Stohlmann [22] focus on determining whether or not a Court's statement invokes the proportionality argument. The task is framed as a binary sentence-level classification. Finally, Habernal et al [8] classify spans of legal text with argument types (e.g., Textual interpretation, Application to the concrete case) and identify the actors responsible for each argumentative content (e.g., ECHR, Applicant, State).

Structures and relations. Beyond identifying discrete components, AM aims to reconstruct the logical architecture that connects such elements. This entails modelling inferential and rhetorical relationships – such as support, rebuttal, or contradiction – between sentences or propositions.

Ali et al. [18] develop methods for support and attack relation classification, combining linguistic cues (e.g., discourse connectors), semantic similarity, and weak supervision strategies They also explore automated dataset construction via weakly supervised methods. Santin et al. [14] expand this space by identifying and classifying five relation types – Support from Premise(s), Support from Failure, Rebuttal, Undercut, and Rephrase – offering a fine-grained typology of argumentative dynamics within judicial texts. This layer of analysis aims to capture the holistic reasoning process of judicial decisions, allowing thee reconstruction of structured argumentative graphs.

Legal reasoning: A more recent evolution in LAM centers on emulating legal reasoning. This shift reflects a broader ambition to go beyond surface-level pattern recognition toward interpreting and replicating the decision-making logic of courts. Chlapanis et al. [9] introduce the Legal Argument Reasoning (LAR) task. Given the case facts, LLMs predict the next logical statement in a legal argument chain, among multiple choice options.

Similarly, Held and Habernal [12] examine whether judges' questions during oral hearings can serve as predictors of subsequent dissenting or concurring opinions. Thus, they frame reasoning as a process of inferring latent intent. Authors examine various tasks including a binary classification task of what kind of opinions are expressed after the questioning part, i.e., "dissenting", "concurring", "partly", "opinion"). Furthermore, reasoning tasks are explored in combination with summarisation and Q&A.

Reasoning is also integrated with other NLP tasks. To evaluate the quality of summaries, Xu et al [30] generate structured question-answer pairs, grounded in the IRC schema. Similarly, Smywinski [31] assesses the capacity of LLMs to reason through question-answer pairs generated from legal texts. Their emphasis is on understanding, interpreting, and reasoning over legal arguments, rather than merely detecting or classifying argument components or their relationships. Lastly, Lu [32] explores prompt engineering as a means of simulating professional legal reasoning, testing whether structured prompts like IRAC or TREACC can elicit logical coherence and improve LLMs' ability to assess legal arguments in a zero-shot setting.

Multi-task and hybrid setups: Some studies approach argument mining through a multi-stage pipeline, combining several sub-tasks to capture argumentative structures more comprehensively. Zhang and others [33] begin with argument clause recognition, classifying sentences from judicial opinions as either argumentative or non-argumentative. The second stage, argument relation mining, identifies inferential links between components within the same argument. Finally, argument component classification assigns each argumentative clause a role — either premise or conclusion—using two binary classifiers. Similarly, Poudyal et al [7] define a three-part pipeline tailored to legal texts: (A) Argument Clause Recognition, which determines whether a clause forms part of an argument; (B) Argument Relation Mining, which assesses whether two clauses are part of the same argument; and (C) Premise/Conclusion Recognition, which classifies the argumentative role of each clause. This integrated approach enables a more structured analysis of legal reasoning across multiple layers.

5. Discussion

Legal argument mining poses several unique challenges for LLMs, rooted in both the nature of legal texts and on the current limitations of computational approaches. A fundamental obstacle is the scarcity of annotated legal data, as highlighted by Zhang et al. [33]. Data scarcity has notoriously been a crucial issue for the argument mining community [2][3]. In the law this issue is aggravated by the complexity and linguistic specificities of legal texts, whose annotation requires costly legal expertise. This aspect makes the task of legal argument mining particularly suitable for experimentation in a zero-shot or few-shot learning setting.

In fact, although domain-specific pre-training of models offers promising solutions to this "data poverty", more research is needed to adapt NLP tools to high-complexity tasks. The issue of label ambiguity, where a clause can function as both a premise and a conclusion, complicates binary classification approaches, as observed by Al-Abdulkarim et al. [28] conducted on ECHR dataset. They also note the brittleness of in-context learning, model performance being sensitive to prompt design. Santin et al. [14] emphasise the difficulty of reconstructing complex argumentative structures due to low annotator agreement and the challenges in long-distance link prediction – especially when argument pairs span large portions of text, increasing class imbalance. Moreover, Habernal et al. [8] argue that there still exists a gap between how arguments are represented in computational argumentation and how legal experts interpret them, according to legal reasoning. In this direction, Bambroo et al. [21] draw attention to the long, unstructured nature of legal documents and the crucial need for explainability in legal AI, where trust hinges on the reference of authoritative legal sources, and the heavy dependence of performance on embedding quality.

Taken together, these challenges remark the need for more robust, interpretable, and domain-aware models tailored to the legal domain. So far, the performance of LLMs for legal argument mining tasks is promising, but there is a large margin for improvement. For example, Al-Abdulkarim et al. [28] found that law-specific models like legal-BERT outperformed general-purpose models such as GPT-3.5 and GPT-4, particularly in identifying conclusions, and local embedding models proved to be competitive alternatives. In other tasks, limitations in LLMs reasoning have emerged. Studies reveal that even advanced prompting techniques, like Chain-of-Thought (CoT), fail to ensure accurate application of legal standards, with LLMs often reverting to outdated doctrines in U.S. federal jurisdiction law. Moreover, models appear to be strongly influenced by the context provided in the introductory sections of datasets [32]. Evaluation methods themselves also warrant scrutiny. For instance, retrieval-based metrics may not adequately capture the qualitative dimensions of legal argument relevance, and the generalizability of findings is limited due to the use of jurisdiction-specific datasets and legal traditions [31].

Finally, efforts to use LLMs for structured reasoning tasks, such as predicting the next step in legal arguments based on ECHR cases, highlight additional limitations [9]. Results are shown to be affected by factors such as dataset construction methods, summarisation of case facts, and the artificial nature of the prediction task itself.

In future work, we plan to extend this survey study by examining in greater detail the methods and models used for legal argument mining tasks, and by broadening the selection of papers to also include those addressing legal argument generation tasks.

Acknowledgments

This work was partially supported by the following projects: CompuLaw - Computable Law - funded by the ERC under the Horizon 2020 (Grant Agreement N. 833647); PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D23005130001); PRIN2022 EQUAL – EQUitableALgorithms (Ref. Prot n. 2022KFLF3E_001 - CUP J53D23005560001); "FAIR - Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", under the European Commission's NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, Partenariato Esteso (PE00000013); TANGO - Grant

Agreement no. 101120763. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to check grammar and spelling. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] K. D. Ashley, Artificial intelligence and legal analytics: new tools for law practice in the digital age, Cambridge University Press, 2017.
[2] J. Lawrence, C. Reed, Argument mining: A survey, Comput. Linguistics 45 (2019) 765–818.
[3] M. Lippi, P. Torroni, Argumentation mining: State of the art and emerging trends, ACM Trans. Internet Techn. 16 (2016) 10:1–10:25.
[4] K. D. Ashley, Automatically extracting meaning from legal texts: opportunities and challenges, Ga. St. UL Rev. 35 (2018) 1117.
[5] R. Mochales, M.-F. Moens, Argumentation mining, Artificial intelligence and law 19 (2011) 1–22.
[6] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, Transactions on Machine Learning Research (2022).
[7] P. Poudyal, J. Šavelka, A. Ieven, M. F. Moens, T. Goncalves, P. Quaresma, ECHR: Legal corpus for argument mining, in: Proceedings of the 7th Workshop on Argument Mining, 2020, pp. 67–75.
[8] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. Spiecker genannt Döhmann, C. Burchard, Mining legal arguments in court decisions, Artificial Intelligence and Law 32 (2024) 1–38.
[9] O. S. Chlapanis, D. Galanis, I. Androutsopoulos, LAR-ECHR: A new legal argument reasoning task and dataset for cases of the european court of human rights, in: Proceedings of the Natural Legal Language Processing Workshop, 2024, pp. 267–279.
[10] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, P. Malakasiotis, Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 226–241.
[11] T. Santosh, R. G. Haddad, M. Grabmair, ECtHR-PCR: A dataset for precedent understanding and prior case retrieval in the european court of human rights (2024).
[12] L. Held, I. Habernal, Lacour!: enabling research on argumentation in hearings of the european court of human rights, Artificial Intelligence and Law (2024) 1–24.
[13] G. Grundler, P. Santin, A. Galassi, F. Galli, F. Godano, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Detecting arguments in CJEU decisions on fiscal state aid, in: Proceedings of the 9th Workshop on Argument Mining, 2022, pp. 143–157.
[14] P. Santin, G. Grundler, A. Galassi, F. Galli, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Argumentation structure prediction in CJEU decisions on fiscal state aid, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, 2023, pp. 247–256.
[15] V. Walker, D. Foerster, J. M. Ponce, M. Rosen, Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment, in: Proceedings of the 5th Workshop on Argument Mining, 2018, pp. 68–78.

[16] H. Xu, K. Ashley, Multi-granularity argument mining in legal texts, in: Legal Knowledge and Information Systems, IOS Press, 2022, pp. 261–266.
[17] H. Xu, J. Savelka, K. D. Ashley, Toward summarizing case decisions via extracting argument issues, reasons, and conclusions, in: Proceedings of the eighteenth international conference on artificial intelligence and law, 2021, pp. 250–254.
[18] B. Ali, S. Pawar, G. Palshikar, R. Singh, Constructing a dataset of support and attack relations in legal arguments in court judgements using linguistic rules, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 491–500.
[19] B. Ali, S. Pawar, G. Palshikar, A. S. Banerjee, D. Singh, Legal argument extraction from court judgements using integer linear programming, in: Proceedings of the 10th Workshop on Argument Mining, 2023, pp. 52–63.
[20] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, A. Wyner, Identification of rhetorical roles of sentences in Indian legal judgments, in: Legal knowledge and information systems, IOS Press, 2019, pp. 3–12.
[21] P. Bambroo, S. Adhikary, P. Bhattacharya, A. Chakraborty, S. Ghosh, K. Ghosh, MARRO: multi-headed attention for rhetorical role labeling in legal documents, Artificial Intelligence and Law (2025) 1–30.
[22] K. Lüders, B. Stohlmann, Classifying proportionality-identification of a legal argument, Artificial Intelligence and Law (2024) 1–28.
[23] G. Grundler, A. Galassi, P. Santin, A. Fidelangeli, F. Galli, E. Palmieri, F. Lagioia, G. Sartor, P. Torroni, et al., AMELIA-Argument Mining Evaluation on Legal documents in ItAlian: A CALAMITA challenge, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, 2024.
[24] J. F. Muñoz-Soro, R. del Hoyo Alonso, R. Montañes, F. Lacueva, A neural network to identify requests, decisions, and arguments in court rulings on custody, Artificial Intelligence and Law 33 (2025) 101–135.
[25] E. Cabrio, S. Villata, Five years of argument mining: a data-driven analysis, in: IJCAI, ijcai.org, 2018, pp. 5427–5433.
[26] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, N. Slonim, Stance classification of context-dependent claims, in: EACL (1), Association for Computational Linguistics, 2017, pp. 251–261.
[27] V. Niculae, J. Park, C. Cardie, Argument mining with structured SVMs and RNNs, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 985–995.
[28] A. Al Zubaer, M. Granitzer, J. Mitrović, Performance analysis of large language models in the domain of legal argument mining, Frontiers in artificial intelligence 6 (2023) 1278796.
[29] M. Elaraby, D. Litman, ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022.
[30] H. Xu, K. Ashley, A question-answering approach to evaluating legal summaries, in: Legal Knowledge and Information Systems, IOS Press, 2023, pp. 293–298.
[31] A. Smywiński-Pohl, T. Libal, Enhancing legal argument retrieval with optimized language model techniques, in: JSAI International Symposium on Artificial Intelligence, Springer, 2024, pp. 93–108.
[32] Y.-A. Lu, H.-Y. Kao, 0x. yuan at semeval-2024 task 5: Enhancing legal argument reasoning with structured prompts, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 385–390.
[33] G. Zhang, D. Lillis, P. Nulty, Can domain pre-training help interdisciplinary researchers from data annotation poverty? A case study of legal argument mining with bert-based transformers, in: Proceedings of the Workshop on Natural Language Processing for Digital Humanities, 2021, pp. 121–130.
[34] O. Shulayeva, A. Siddharthan, A. Wyner, Recognizing cited facts and principles in legal judgements, Artificial Intelligence and Law 25 (2017) 107–126.

Appendix

A. Overview of legal argument mining datasets

Dataset	Authors	Court	Area of law	Language	No. documents	Timespan	Labels	Statistics	Annotation	Tool / Format	Link
European Courts											
ECHR	Poudyal, et al. [7]	ECHR	Fundamental rights	English	42 cases: 20 decisions, 22 judgments	N/S	Premise, conclusion, and non-argument, along with links between premises and conclusions	1951 premises, 743 conclusions	2 + 2 legal expert annota- tors. Cohen's k=0.80 (final agreement)	Gloss	http://www.di.uevora.pt/~pq/echr/
LAM-ECHR	Habernal, et al. [8]	ECHR	Fundamental rights (Art 3,7,8 EU Charter)	English	373 judgments	N/S	5 actors: ECHR, Applicant, State, Third Parties, Commission/ Chamber. 16 argument types: incl. procedural arguments & methods of interpretation (textual, historical, teleological, etc.)	2.3M tokens; 15.205 argument spans	6 law students (supervised by 2 postdocs, consulted by 2 law professors). Krippendorff's unitised α>0.80 depending on the label	UIMA-XMI stand-off annotation format	https://github.com/trusthlt/mining-legal-arguments
LAR-ECHR	Chlapanis, et al. [9]	ECHR	Fundamental rights	English	191 cases	N/S	Target argument criteria: authored by judges; classified under 'Application to the concrete case'; and must be the first such argument after the parties' submissions. Distractor selection: drawn from the same corpus, chosen to match the target in style and vocabulary, and must not be paraphrases of the target	403 samples/ instances. Split into few-shot (5), development (98), and test (300) sets	no human annotators	N/S	https://huggingface.co/datasets/AUEB-NLP/lar-echr
Demo-sthenes	Grundler, et al. [13]	CJEU	Fiscal state aid	English	40 judgments	2000-2018	Focus on 'Findings of the Court' section. Hierarchical annotation scheme: (1) argument- ative components (premises, conclusions), (2) premise types (legal or factual), and (3) argumentation schemes (e.g., Rule, Precedent, Authority)	9320 sentences: 2375 premises, 160 conclusions, 1575 factual, 906 legal. Schemes: 53 aut, 56 class, 296 itpr, 503 prec, 15 princ, 322 rule	two legal experts; agreements: premises/conclu- sions Cohen's k=0.95, factual=0.87 legal=0.82, schemes from 0.43 (princ) to 0.93 (rule)	XML	https://github.com/adele-project/demosthenes
Demo-sthenes (ext.)	Santin, et al. [14]						Extends the original Demosthenes dataset with relational labels between components: support, rebut, undercut				
LaCour!	Held & Habernal [12]	ECHR	Fundamental rights	English and French	154 transcribed ECHR court- room dialogues	2012-2021	Sentence-level labels: questions, opinions (e.g., dissenting, con- curring), speaker roles, language and time- stamps. Each hearing is linked to the corres- ponding final judgment	267 hours of courtroom dialogue. 2.16 million word tokens. On average, 14,038 words and 557 sentences per hearing.	The transcripts are partially manually corrected and annotated with speaker roles and language labels	N/S	https://huggingface.co/datasets/TrustHLT/LaCour
National courts - civil law											

Italian Dataset AMELIA	Grundler, et al. [23]	Regio- nal Tax Comm.	Tax law	Italian	225 VAT rulings	2010 - 2022	Three-level hierarchical annotation: argument component, premise type, argument schemes (as defined for the Demosthenes corpus)	(1) argument components: prem (2910), concl(401); (2) types: facts (1892), legal(1338) (3) schemes: 144 princ, 100 class, 374 itpr, 419 prec, 118 rule. The splitting between train, validation, and test data was done at the document level (60:20:20)	4 tax law experts	XML	https://github.com/adele-project/AMELIA/
Spanish Dataset	Muñoz-Soro, et al.[24]	Provin- cial and higher courts	Family law	Spanish	3,047 court rulings, sourced from CENDOJ	2006-2020 (priority given to more re- cent cases)	Types of requests (custody, assigned family home, alimony); decision; judicial principles; justifications (legal, factual)	Subset of 2394 child custody rulings, 36,087 labels (mean: 15.07 labels per case). Dataset split: training data (72%), test data (18%), and validation data (10%)	2 legal experts (1 PhD, 1 LLM); annotator agree- ment: F1 varies from 0.60 to 0.86, Kappa index: varies from 0.33 to 0.73	BRAT tool [26]	https://github.com/labje/bidaraciv
German Dataset	Lüders & Stohl- mann [22]	German Federal Const. Court (GFCC)	Constitutional law	German	300 randomly selected decisions, 245 of which contained relevant merits sections	1951 - 2021	Annotations, limited to the "merit" sections, are based on the GFCC's four-step proportio- nality test (legitimate aim, suitability, ne- cessity, balancing) and allows multiple labels per sentence	54,929 sentences, out of which 24,577 sentences were deemed relevant after filtering for merits	13 legally trai- ned annotators, three iterative cycles, reaching a final Fleiss kappa of 0.78	N/S	https://zenodo.org/records/10513684
National courts - common law											
Indian Supreme Court corpus	Ali, et al. [18]	Supr. Court of India	N/S	English	30,034 decisions	1952 - 2012	Support and attack relations	4,062,500 sentences out of which 20,506 sentence pairs were identified as containing Support (10,746) or Attack (9,760) relations	The relational sentence pairs were automa- tically identified and labelled. An estimated over- all precision of 71.6 %	N/S	The dataset would be shared upon request
Supreme Court of India (ext.)	Ali, et al. [19]	Supr. Court of India	Industrial disputes	English	10 judgements	N/S	Each argument is rep- resented by its start and end sentence num- bers where each sen- tence in between is considered as a part of the argument, addit- ionally annotated for major claims within the arguments	1524 sentences spread across 418 paragraphs out of which 127 were identified as gold standard arguments	Annotators not specified; F1 scores for the agreement: Arg-exact=0.3, Arg-subset=0.47 Arg-overlap=0.56, and Arg-sentences=0.59.	N/S	The dataset would be shared upon request
DIN + DUK Dataset	Bambroo, et al. [21]	DIN: Indian Supr. Court; DUK: UK Supr. Court	5 legal do- mains: crimi- nal, land/ property, constitutional, labour/ indus- trial, and IP	English	DIN: 150 judgments (100 newly annotated and 50 from prior work); DUK: 50 judgments (pre- viously released in earlier re- search)	N/S	Each sentence labeled using a seven-role schema – Facts, Lower Court Ruling, Argu- ment, Ratio, Statute, Precedent, Present Court Ruling – that reflects the internal structure of legal decisions	DIN: 30,729 sentences; DUK: 18,155 sentences	Law graduates. Agreement F1-scores: 0.921 (DIN) and 0.915 (DUK)	GATE3 Tool [34]	https://github.com/purbid/MARRO_Rhetorical-Role-Labeling
BVA dataset	Walker, et al. [15]	N/S	Veteran PTSD-claims	English	30 Veteran appeal decisions	2013-2016	Sentences annotated with 3 rhetorical roles, relevant to evidentiary reasoning: evidence, reasoning, and findings — mapping directly to how tribunals interpret and assess claims	8149 sentences: 1412 evidence, 422 reasoning, 310 findings sentence	2 legal experts. Type of evi- dence: k= 0.76, Credibility factor: k= 0.66, Pattern/Soft Rule: k= 0.53, Decision: k= 0.80	N/S	N/S

Canada dataset	Xu, et al. [17] [16] [30]	N/S	N/S	English	28,733 cases and their corres- ponding summaries, all sourced from the CanLII. 574 randomly selec- ted pairs from the 28,733 case/summary pairs. Later extended to 1049	N/S	Issue, conclusion, and reason-sentences (IRC)	The total number of sentences from the correspond- ing full texts is 120,707, which is significantly more than the summaries' 7,484 sentences	2 law students. The mean of Cohen's k co- efficients across all types for summaries =0.734; the mean for full case texts=0.602	N/S	N/S

B. Overview of argument mining tasks and models

Authors	Dataset	Task description	Models
		Detection and Classification	
Al Zubaer, et al. [28]	ECHR	Argument clause classification: binary, conclusion/non-conculsion, premise/non-premise	GPT-3.5-turbo, GPT-4, OpenAssistant LLaMA (instruction-fine-tuned on LLaMA-1), RoBERTA, Legal-BERT
Ali et al. [19]	Indian Supreme Court	Text segmentation using argument markers. Assigning sentence-level binary labels for argument markers, i.e. claim, argumentative, etc. Output: structured sentence span, i.e. extracted argument. These local classifications are then integrated using Integer Linear Programming (ILP) to produce an optimal, document-level segmentation of legal arguments	RoBERTA-large
Bambroo et al. [21]	DIN + DUK Dataset	Multi-class sentence classification task into rhetorical roles, i.e. facts, ratio, precedent etc.	MARRO, Hierarchical BiLSTM-CRF, LEGAL-BERT-BiLSTM-CRF, sciBERT-HSLN, MTL (BERT-SC), Gemini-pro
Elaraby & Litman [29]	(reuse) Canada dataset	Multi-class (four classes) sentence-level class- ification. Each sentence is classified into one of three legal argument roles, i.e. issue- reason- conclusion (IRC framework), plus non-argumentative sentence	BERT, BART, Longformer
Grundler et al. [23]	Italian Dataset (AMELIA)	(1) Classifying an argumentative component as either a premise or a conclusion, binary classification, (2) Premise type classification, determining whether a given premise is factual, legal, or both. Multi-label binary classification task, (3) Argument scheme classification, classifying legal premises according to one or more argumentative schemes they instantiate - multi-label multi-class classification	N/S
Grundler et al. [13]	Demosthenes	(1) Argument Detection (AD): given a sentence, classify it as premise, conclusion, or neither; (2) Argument Classification (AC): given a sentence that is known to be argumentative, classify it as premise or conclusion; (3) Type Classification (TC): a multi-label classification problem where a sentence that is known to be a premise is classified as legal (L) and/or factual (F);(4) Scheme Classification (SC): a multi-label classification task where a sentence, known to be a legal premise, is classified according to its scheme; due to the low number of samples in the dataset, the Princ scheme has not been considered	Linear SVC, SVC, Random Forest, Gaussian Naive Bayes and K-Neighbours
Lüders & Stohlmann [22]	German Dataset	Binary sentence-level classification task, determining whether a sentence from the GFCC decision invokes the argument of proportionality or not	SVC, BERT
Muñoz-Soro, et al. [24]	Spanish Dataset	(1) binary sentence classification, i.e. argumentative/ non-argumentative; (2) Multi-label classification to identify the content of each sentence, i.e. requests, judicial decisions, supporting arguments	BERT-base-multilingual-uncased, DistilBERT-base-multilingual-uncased, XLM-mlm-enro-1024, and XLM-RoBERTa-large, BETO
Walker et al. [15]	BVA dataset	Multi-class single label classification of sentences. Three categories for the classification task, i.e. evidence sentence, reasoning sentence, conclusion sentence on whether a condition of a legal rule has been met	N/S
Xu et al. [16]	(reuse) Canada dataset	refine the IRC approach, by incorporating a token-level and subsequent sentence-level classification of (i) court-issues; (ii) conclusions on such issues (court's decision); and (iii) the court's reasons for so concluding, coupling the binary classification task with abstractive summarisation (all unannotated sentence are treated as non-IRC sentences)	Legal-BERT, BERT, Longformer
Xu et al. [17]	Canada dataset	Multi-class sentence classification task following the IRC framework, i.e. issue, reason, conclusion, non-IRC	LSTM, CNN, RoBERTa, CNN-BERT

Habernal et al. [8]	LAM:ECHR	(1) Multi-class sequence labeling task. Classifying spans (token-level sequences) according to two: argument type and argument actor, two single-label classification tasks. Each token within a paragraph aph is labelled as to indicate whether (i) it begins (ii) is inside (iii) is outside a given argument span of a certain type. (2) Argument actor classification task, multi-class, single-label classification. Tokens are labeled for their rhetorical or legal role and the actor	RoBERTa-Large, Legal-BERT, SVM

Structures and Relations

Ali et al. [18]	Indian Supreme Court corpus	Multi-class classification task, predict the correct relation label for any given sentence pair. i.e. support, attack, no relation	BERT
Santin et al. [14]	Demosthenes	Binary classification task and as an inferential relation prediction, i.e. support, rebuttal, undercut, rephrase. Link prediction between argumentative components within judicial decisions	ResAttArg, DistilRoBERTa

Legal Reasoning

Chlapanis et al. [9]	LAR-ECHR	Multiple-choice next-statement selection task, forced-choice classification task. Legal Argument Reasoning (LAR): given a sentence, predict the next logical argument statement among multiple possible choices	GPT-4o (L), GPT-4o-mini (S), Mistral-8x22B (L), Mistral-8x7B (M), Mistral-7B (S), Llama-3.1-70B (L), Llama-3.1-8B (S). L, M, S denote the largest, medium, smallest models per family, respectively
Held & Habernal [12]	LaCour!	Multi-class and binary classification task to investigate whether the questions posed by judges during ECHR oral hearings correlate with the type of opinion that those judges later issue in the final judgment, i.e. dissenting, concurring, or none	BERT, Legal-BERT, Legal-RoBERTa, RoBERTa-Large, Llama-3 8b
Lu et al. [32]	(reuse SemEval-2024 Task 5, domain of U.S. civil procedure)	Binary classification task to evaluate whether structured legal reasoning prompts (e.g. IRAC, TREACC) can guide LLMs to determine whether argumentative legal answers are correct or incorrect, based on the context of U.S. civil procedure cases, including the legal question and explanation, in a zero-shot learning setting	Mixtral-8x7B
Xu et al. [30]	(reuse) Canada dataset	Evaluate the presence of argumentative structure following the IRC framework (Issue- Reason-Conclusion), within legal summaries, through a question-answering framework	GPT-4, Longformer Encoder-Decoder (LED), BART
Smywinski et al. [31]	N/S	Improving the retrieval of legally relevant arguments to extract and rank legal arguments from case law in response to a legal query	DeBERTa v.3, Legal-BERT, RoBERTa

Multi-task / Hybrid

Poudyal et al. [7]	ECHR	(1) Argument clause recognition, binary classifi- cation to determine whether a clause is argument- ative or non-argumentative. (2) Premise and conclusion recognition, binary classification, given a set of previously identified argumentative clauses, assign each one either a premise or conclusion label. (3) Argument relation mining, binary class- ification, given a pair of argumentative clauses, classify whether they are part of the same argument structure or not	RoBERTa
Zhang et al. [33]	(reuse) ECHR	(1) Argument clause recognition, binary sentence classification task, i.e. argument clause/ non-argument clause; (2) Argument component classification, two separate binary classification tasks (i) whether an argument clause is a premise (ii) whether an argument clause is a conclusion. A clause can be both a premise and a conclusion in different arguments. (3) Argument relation mining, binary classification, each pair of clauses is classified as either related (i.e. belonging to the same argument structure) or not related	RoBERTa, one-layer BiLSTM, Legal-BERT, C-Legal-BERT