An LLM-based Tool for Legal Argument Annotation

Karla Salas-Jimenez[1,*], Jean-Guy Mailly[1], Leila Moudjari[1] and Laurent Perrussel[1]

[1]IRIT, Université Toulouse Capitole, France

Abstract

We present a Web interface that we developed for using an LLM-based approach that identifies arguments in legal texts. Our tool will facilitate the annotation of arguments by legal scholars or practitioners, thus providing new datasets for legal argument mining tasks.

Keywords

Natural Argument Annotation, Legal Argumentation, Large Language Models, User Interface

1. Introduction

Although argument mining [1, 2] has been a very active field, there is a surprisingly small number of annotated datasets in the domain of legal argumentation. While some recent works have proposed legal argument datasets (*e.g.* notably from the European Court of Human Rights [3] or the US Supreme Court [4]), it is noticed that the number and variety of such datasets remains limited [5]. Moreover, given the nature of legal argumentation compared to other forms of argumentation (*e.g.* political debates or social networks discussions), it is important to train argument mining systems with suitable corpora to improve their performance in the legal field, arbitrary argumentative data may not be well-suited for this task.

For this reason, we have started evaluating an LLM-based methodology for identifying the arguments in legal texts [6, 7]. Similar LLM-based approaches for data annotation have been proved successful for annotating arguments in other domains than Law [8], as well as for other tasks [9, 10]. Our aim is not to entirely replace argument mining by LLMs, but to assist human annotators with LLMs in order to facilitate the creation of new legal arguments copora.

2. Our Tool LexAgree

Our system is a web-based application that uses LLMs to assist users in annotating arguments in legal texts. The underlying approach for identifying arguments is a significantly improved version of the pipeline proposed by [6, 7].[1] The system takes a legal text as input and automatically identifies and visualizes the arguments contained in the document. As shown in Figure 1, it consists of three screens. In the first screen, users can upload a legal document in `.txt` format and select the LLMs to be used for argument extraction. Currently, the system supports three models: `meta-llama/Meta-Llama-3.1-8B-Instruct`, `Equall/Saul-7B-Instruct-v1`, and `Qwen/Qwen2.5-7B-Instruct`. The selected models independently analyze the document, and their outputs are subsequently processed to obtain the final set of arguments.

The second screen presents the progress of the analysis and provides information about the different steps performed by the system. The processing pipeline consists of the following stages:

Segmenting the document. Because of the input-length limitations of the LLMs, the document is first divided into smaller segments. Rather than splitting the document only according to a fixed number of tokens or sentences, we aim to identify points at which the topic of the text changes. For

CMNA'26: 26th International Workshop on Computational Models of Natural Argument, September 14, 2026, Barcelona, Spain

*Corresponding author.

Karla-Denia.Salas-Jimenez@irit.fr (K. Salas-Jimenez); jean-guy.mailly@irit.fr (J. Mailly); leila.moudjari@irit.fr (L. Moudjari); laurent.perrussel@ut-capitole.fr (L. Perrussel)

[1]A detailed presentation of this approach is left for future publication and is beyond the scope of the present demo paper.

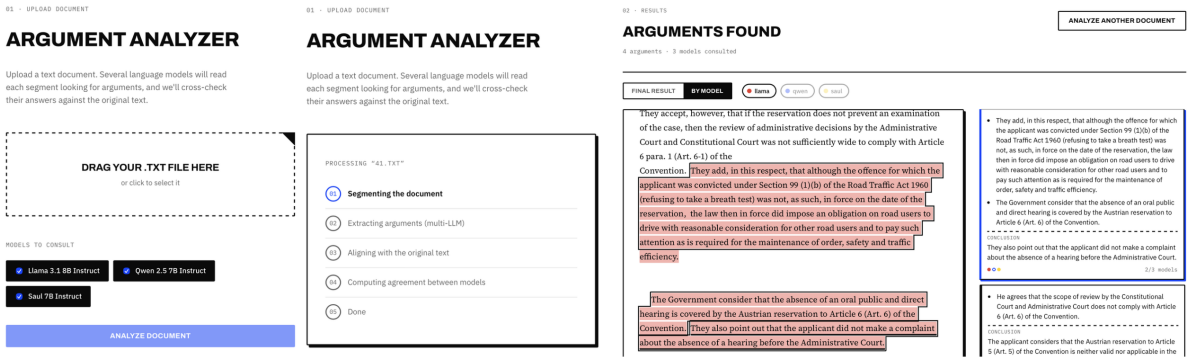

Figure 1: Overview of the LexAgree Web interface: (a) document upload and LLM selection, (b) argument extraction pipeline and progress, and (c) visualization and inspection of the extracted arguments.

this purpose, we follow approaches similar to those proposed by [11] and merge consecutive segments when they maintain topic continuity. Topic shifts are identified using sentence embeddings and cosine similarity, following ideas inspired by [12, 13]. In the current implementation, this stage also includes a preprocessing step that retains the sections of ECHR documents that are more likely to contain argumentative content.

Extracting arguments. In this stage, each selected LLM is prompted to identify the arguments contained in the corresponding document segments. All selected models receive the same prompt and are asked to perform the same extraction task. The prompt follows the approach used for Pipeline 1 proposed by [6, 7].

Aligning with the original text. The arguments generated by the LLMs do not necessarily reproduce the original text verbatim. In some cases, a model may paraphrase an argument while preserving its meaning. To ensure that the extracted arguments can be located and visualized in the original document, we therefore include an alignment step that matches the generated arguments with their corresponding spans in the source text.

Computing agreement between models. This stage is motivated by the idea that multiple annotators are generally required when creating reliable annotations, while also reducing the number of arguments that are supported by only a single LLM. We group similar arguments across models and construct a final argument from each group. For each group, the most frequently occurring claim is selected as the final claim. Among similar premises, we prioritize those containing more specific information, such as references to articles, and those providing more information in the form of longer text spans. The resulting set of arguments represents the final output of the agreement stage.

The third screen provides several complementary ways of inspecting the extracted arguments. Users can either visualize the final arguments obtained after the agreement stage or inspect the arguments produced by individual LLMs. In both cases, the interface provides two complementary views. First, the original legal text is displayed with the detected argument spans highlighted, allowing users to inspect the arguments in their original context. Second, the extracted arguments are displayed separately, distinguishing between their premises and conclusion and indicating the level of agreement among the selected models. This allows users to compare the individual model outputs with the aggregated result and to inspect how consistently the models identified each argument.

The source code of the system is publicly available online[2] as well as a demonstration video[3].

As future work we plan to extend LexAgree in two directions: (i) supporting the annotation of new datasets through argument evaluation and ranking, using the resulting feedback to assess its usefulness for human annotation; and (ii) evaluating the system beyond ECHR documents and `.txt` inputs by automatically identifying argument-relevant sections instead of relying on document-specific structure.

[2] https://github.com/KarlaDSJ/LexAgree
[3] https://github.com/KarlaDSJ/LexAgree/blob/main/Demo%20LexAgree.mov

Acknowledgments

This work is funded by the French National Research Agency under the grant ANR-22-CPJ1-0061-01.

Declaration on Generative AI

The authors have not employed any Generative AI tools for writing this article. Claude was used for designing the Web interface.

References

[1] E. Cabrio, S. Villata, Five years of argument mining: a data-driven analysis, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 5427–5433. URL: https://doi.org/10.24963/ijcai.2018/766. doi:10.24963/IJCAI.2018/766.

[2] M. Lenz, R. Bergmann, User-centric argument mining with arguemapper and arguebuf, in: Computational Models of Argument - Proceedings of COMMA 2022, Cardiff, Wales, UK, 14-16 September 2022, volume 353 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2022, pp. 367–368.

[3] P. Poudyal, J. Savelka, A. Ieven, M. F. Moens, T. Goncalves, P. Quaresma, ECHR: Legal corpus for argument mining, in: Proceedings of the 7th Workshop on Argument Mining, Association for Computational Linguistics, Online, 2020, pp. 67–75.

[4] S. Wang, L. Pobbathi, H. Chen, LAMUS: A large-scale corpus for legal argument mining from U.S. caselaw using llms, CoRR abs/2603.08286 (2026). URL: https://doi.org/10.48550/arXiv.2603.08286. doi:10.48550/ARXIV.2603.08286. arXiv:2603.08286.

[5] R. Liepina, F. Galloni, F. Lagioia, M. Lippi, M. Musicco, B. Sayin, A. Passerini, G. Sartor, Legal argument mining: Recent trends and open challenges, in: T. Koref, L. Held, I. Habernal (Eds.), Proceedings of the First Argument Mining and Empirical Legal Research Workshop co-located with the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025), Chicago, United States, June 20, 2025, volume 4089 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1–14. URL: https://ceur-ws.org/Vol-4089/paper1.pdf.

[6] C. Berghegger, C. Philippe, K. Salas-Jimenez, J.-G. Mailly, L. Moudjari, L. Perrussel, Discovering the potential of llms in annotating legal texts for argument mining, in: O. Cocarascu, S. Doutre, J.-G. Mailly, A. Rago (Eds.), Proceedings of the Second International Workshop on Argumentation and Applications (Arg&App 2025), Melbourne, Australia, November 11, 2025, 2025, pp. 33–44. URL: https://hal.science/hal-05339535/document#page=35.

[7] C. Berghegger, C. Philippe, K. Salas-Jimenez, J.-G. Mailly, L. Moudjari, L. Perrussel, Discovering the potential of llms in annotating legal texts for argument mining (extended abstract), in: R. Markovich, L. D. Caro, A. Rapp, C. Schifanella (Eds.), Legal Knowledge and Information Systems - JURIX 2025: The Thirty-eighth Annual Conference, Turin, Italy, 9-11 December 2025, volume 416 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2025, pp. 382–384. URL: https://doi.org/10.3233/FAIA251615. doi:10.3233/FAIA251615.

[8] A. Lindahl, LLMs as annotators of argumentation, in: L. Frermann, M. Stevenson (Eds.), Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025), Association for Computational Linguistics, Suzhou, China, 2025, pp. 242–252. URL: https://aclanthology.org/2025.starsem-1.19/. doi:10.18653/v1/2025.starsem-1.19.

[9] M. A. Gray, J. Savelka, W. M. Oliver, K. D. Ashley, Can GPT alleviate the burden of annotation?, in: G. Sileno, J. Spanakis, G. van Dijck (Eds.), Legal Knowledge and Information Systems - JURIX 2023: The Thirty-sixth Annual Conference, Maastricht, The Netherlands, 18-20 December 2023, volume 379 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2023, pp. 157–166. URL: https://doi.org/10.3233/FAIA230961. doi:10.3233/FAIA230961.

[10] J. Savelka, K. D. Ashley, The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts, Frontiers Artif. Intell. 6 (2023). URL: https://doi.org/10.3389/frai.2023.1279794. doi:10.3389/FRAI.2023.1279794.

[11] M. Frohmann, I. Sterner, I. Vulić, B. Minixhofer, M. Schedl, Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11908–11941. URL: https://aclanthology.org/2024.emnlp-main.665/. doi:10.18653/v1/2024.emnlp-main.665.

[12] A. Krassovitskiy, R. Mussabayev, K. Yakunin, Llm-enhanced semantic text segmentation, Applied Sciences 15 (2025). URL: https://www.mdpi.com/2076-3417/15/19/10849. doi:10.3390/app151910849.

[13] T. Brugger, M. Stürmer, J. Niklaus, Multilegalsbd: A multilingual legal sentence boundary detection dataset, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 42–51. URL: https://doi.org/10.1145/3594536.3595132. doi:10.1145/3594536.3595132.