From Construction to Application

Sahitaj, Premtim; Ruiz-Dolz, Ramon; Sahitaj, Ariana; Nizamoglu, Ata; Schmitt, Vera; Mohtaj, Salar

Published in:
Computational Models of Argument

DOI:
10.3233/FAIA240324

Publication date:
2024

Licence:
CC BY-NC

Document Version
Publisher's PDF, also known as Version of record

Computational Models of Argument
C. Reed et al. (Eds.)

doi:10.3233/FAIA240324

From Construction to Application: Advancing Argument Mining with the Large-Scale KIALOPRIME Dataset

Premtim SAHITAJ [a,b,1], Ramon RUIZ-DOLZ [c], Ariana SAHITAJ [a], Ata NIZAMOGLU [a,b], Vera SCHMITT [a,b], Salar MOHTAJ [a,b], and Sebastian MÖLLER [a,b]

[a] *Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany*
[b] *German Research Center for Artificial Intelligence (DFKI), Berlin, Germany*
[c] *Centre for Argument Technology, University of Dundee, United Kingdom*

Abstract. In this study, we introduce KIALOPRIME, a novel large-scale dataset comprising 5,687 argument discussion graphs with a total of 1,088,801 of supporting, attacking, and neutral argument relations, derived from the structured debates of the online discussion platform Kialo.com. This dataset facilitates in-depth analysis of argument structures and the dynamics of discourse, serving as a substantial resource for computational argumentation research. We explore argument inference through traditional sequence classification and a modern generative reasoning based approach, employing an open-source mixture of experts LLM to interpret and enrich each argument pair with high-quality synthetic elaborations about the argumentative interaction. We achieve baseline results of F1 .899 and .840 within discussions and F1 .908 and .840 across discussions for the argument relation and elaboration classification models, respectively. While the elaboration-based model scores slightly lower on the classification task, we highlight areas of improvement to better capture the hidden complexities of argumentative text. These initial findings are promising as they not only establish robust benchmarks for future studies but also demonstrate the potential for using generative reasoning to provide a more insightful analysis of argument relations.

Keywords. argument mining, argument inference, large language models

1. Introduction

Argument mining has been established as an important area of application within the domain of computational argumentation by leveraging natural language processing (NLP) techniques to automatically identify and extract the argumentative structure present in unstructured texts. When we identify the components of argumentation, we are able to understand the positions that are being held and the reasons for holding them [1]. By parsing through large volumes of textual data, argument mining systems provide access to insights from various areas of public opinion such as social media [2], online debate

[1] Corresponding Author: Premtim Sahitaj, sahitaj@tu-berlin.de

[3], or news editorials [4]. The main identified limitations of argument mining systems reside in the size of publicly available corpora [5,6], the domains or topics represented in these corpora [7], and the lack of transparency of such systems [8]. Therefore, advancing in these aspects will result in significant contributions to the work in computational argumentation, specifically in argument mining.

In this paper, we introduce a large-scale dataset, KIALO-PRIME, comprising 5,687 argument graphs and 1,088,801 argument pairs sourced from the moderated debate platform Kialo[2]. The argument pairs are labeled as a directed inference from one argumentative discourse unit (ADU) as evidence or premise to the other as a claim or target. The directed inference is annotated as either support, attack, or neutral. The purpose of KIALO-PRIME is to accelerate research in computational argumentation by providing a robust source for investigating structured argumentation across various topics and domains. We exemplify the many opportunities that KIALOPRIME provides for the research community by employing advanced methodologies, including using a mixture of experts Large Language Model (LLM) to generate synthetic elaborations of argumentative interactions. To the best of our knowledge, the presented dataset is the largest available and the most comprehensive collection of Kialo argument discussion graphs. A modified version of the dataset is available for the non-profit research community on request.[3]

2. Related Work

Sometimes, it can be difficult to define a unique and standard way of approaching argument mining, but most of the work done in this area can be grouped into three major tasks: argument segmentation, argument classification, and argument relation identification [1,9,10,11]. First, argument segmentation aims at identifying the argumentatively relevant spans or ADUs from unstructured text inputs [12]. Second, argument classification focuses on identifying the role of the ADUs and providing additional information, for example, classifying these ADUs as premises or claims depending on their purpose in the argument [13]. Third, argument relation identification aims to structure the previous information by identifying relations and classifying them into one of the standard argumentative relation classes (i.e., attack/conflict, support/inference) [14]. This way, after the whole argument mining process, the original unstructured natural language input acquires a graph structure that highlights most of the relevant argumentative information [15,16]. Figure 1 depicts the resulting structure of two extracted arguments and their identified reference. Argument relation identification and classification represent the most challenging task within argument mining [10]. This is mainly due to the semantic complexity of understanding argumentative relations by just modeling the natural language text included in the ADUs. Additional contextual information about the ADU interaction can be helpful in this task [3,17]. However, the limited availability of large-scale comprehensive datasets containing argumentative information, specifically regarding argument relations, makes it difficult to investigate this direction in-depth. Stab et al. [18] developed a corpus of 25,000 relations across eight controversial topics to evaluate argument mining systems. Each relation is categorized as pro, con, or neutral, and the structure is flat with each argument tied to one of the topics as a major claim. This con-

[2] www.kialo.com

[3] https://github.com/news-polygraph/dataset-kialoprime

trasts with the complex structures of online debate graphs, which represent more diverse perspectives within the same context. Agarwal et al. [19] created a dataset from Kialo with 1,560 discussion graphs, focusing solely on support and attack relations.

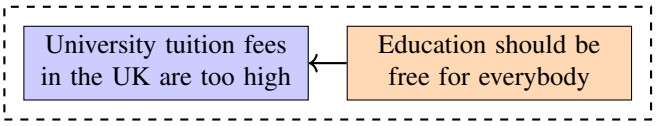

Figure 1. Example premise (orange) referencing a claim (blue).

Therefore, human annotation is an essential part of argument mining, providing ground truth data for computational models to learn from by identifying ADUs and categorizing their interactions [20,21]. Nevertheless, this method faces obvious challenges of scalability and consistency due to its cost-intensive nature and susceptibility to subjective bias, particularly in large or complex datasets [22]. The resulting inconsistencies must be addressed to ensure the reliability of the argument mining systems. To mitigate these challenges, researchers have proposed various solutions. One approach is to employ multiple annotators and use statistical measures like Kappa scores to evaluate inter-annotator agreement, thus ensuring more consistent annotations [23]. CASS was defined as a variation of the metric specifically defined for the computational argumentation domain to improve the quality of the inter-annotator agreement calculations by taking into consideration the structural complexity of argument annotations [24]. A different strategy in this direction is the development of semi-automated tools that assist annotators by suggesting potential arguments and their relations based on preliminary machine learning models, reducing the workload and subjective bias [16,25].
The recent advances in decoder-based LLMs offer a promising solution to some of these limitations. Trained on diverse and increasingly large datasets, LLMs excel in text-to-text problem settings, facilitating the automation of information extraction tasks in various domains without needing large training datasets [26]. Increased context sizes of up to 128k tokens [27] and the integration of external knowledge bases enable guided generation of text [28,29]. Despite challenges in evaluating the quality of LLM outputs, highly specific benchmark results are promising [30]. For example, the argumentation domain [25] shows that LLMs can produce coherent, contextually relevant text on specific tasks and situations.

3. Corpus Creation

Online discussion platforms offer a great opportunity to analyze argumentation due to their structured format and quality content. In contrast to flat argumentative structures [20,18], these platforms provide complex hierarchies of argumentation. In this dialogical setting, participants dynamically integrate opinions as the discussion progresses over time. Often, we can observe that arguments in the upper levels of the discussion facilitate broader discourse on the general topic, while arguments on the deeper levels present more specialized argumentation. For instance, a discussion starting with the major claim *"Does God exist?"* can evolve from broad philosophical debates to specific (fringe) argumentation.

3.1. Collection

For our research, we chose Kialo as the primary source for creating a large-scale argument graph dataset due to its structured debate format and the availability of discussion-level metadata. Kialo's detailed, graph-based model of argumentation facilitates the extensive analysis of argument relations. Discussions are initiated with a title, a thesis statement that is being discussed, and background information on the topic. The major claim or thesis statement is the root argument from which users are prompted to select whether they would like to add a supporting (pro) or attacking (con) argument to a selected existing argument. Arguments entered into a discussion me be self-contained. Additionally, duplicate hints are presented to ensure that an argument is only added once to a discussion. Due to moderation capabilities, arguments can be removed, edited, or re-structured by a selected set of users. When enabled, users can vote on the impact of each argument, which reflects the argument's veracity and relevance. From Kialo, we collect and process 6,818 discussion graphs. Table 1 details the distribution of languages among the collected discussion graphs, which illustrates the predominance of English on the discussion platform. Even though modern multilingual modeling approaches enable reasoning across a variety of tasks in the presented language set, we concentrate on the English subset of the collected discussions. This leaves us with 5,687 high-quality discussion graphs for further analysis.

Table 1. Language distribution of collected discussion graphs.

Language	Count	Percentage
English	**5,687**	**83.42%**
Italian	429	6.29%
Spanish	142	2.08%
French	131	1.92%
German	123	1.80%
Other	306	4.49%
$\sum$	6,818	100.00%

3.2. Corpus Analysis

From the collected discourse structures, we extract 636,213 individual claims with 725,868 relations, of which 372,501 are supporting, and 353,367 are attacking. Due to the setup of the platform, there are no neutral relations in the dataset. Subsection 3.3 describes how we construct the missing neutral relations and create the three-class dataset that enables us to identify whether an argument pair is related (pro or con) or neutral (non). To illustrate the distribution of topics within the discussion graphs, we utilize a topic modeling strategy based on the discussion graph descriptions. The selected method [31] implements dimensionality reduction on embedding representations and applies density-based clustering to identify clusters of density-connected documents and noise efficiently. With a minimum count of 100 discussion graphs per topic cluster, we automatically identify the largest clusters that account for 4,669 out of 5,687. The remaining discussion graphs were not assigned to density clusters and thus are considered

noise, which should be interpreted as too distant from other discussions in the semantic embedding space. Table 2 illustrates the identified clusters described by representative keywords ranked by count.

Table 2. Topic modeling results of discussion graphs with BERTopic and minimum count of 100.

Count	Topic Clusters
741	Ethics - Law - Society
719	Technology - Science - AI
633	Religion - Philosophy - God
573	Politics - Elections- Democracy
422	Gender - Feminism - Sex - LGBTQ
400	Economics - Capitalism - Politics
399	Entertainment - Sports - Games
310	Climate Change - Environment - Animals
293	Education - School - Programming
179	Social Media - Media - News

3.3. Constructing Neutral Relations

In argument mining settings, we are usually presented with unstructured text and are tasked with extracting argumentative structures. Thus, defining the task of argumentative relation classification with binary targets as either support or attack may be overly restrictive for downstream applications, as the assumption of prior knowledge about argumentative relevance often does not apply. Introducing a neutral category into the classification scheme of argument relations can provide a more nuanced understanding of argument dynamics, recognizing that not all discourse interactions result in clear supportive or contradicting relations. We construct neutral relations by random-sampling claims of argument pairs that not part of the same discussion graph, assuming these pairs are unlikely to have a direct argumentative relation. Consequently, we are left with non-argumentative $362,933$ argument pairs. Table 3 describes the final counts of cases per label. This approach enriches our training data, allowing downstream models to distinguish between related (pro and con) and unrelated (non) argument pairs, enhancing the utility of the model for real-world applications where neutral interactions are common.

Table 3. Distribution of classes

Relation	Count	Percentage
Support	372,501	34.21%
Attack	353,367	33.33%
Neutral	362,933	32.45%
$\sum$	1,088,801	100.00%

3.4. Elaborating on Argument Relations

In the field of argument mining, we traditionally adopt transformer-based language models due to their high effectiveness across a variety of tasks and domains [16,32]. However, these models generally act as black boxes, which poses significant challenges in terms of interpretability and transparency. In conventional setups, the challenge of identifying argument relations is often formulated as either a binary or tertiary sequence classification learning problem. Often, these labels are annotated manually by a group of annotators where subjective tendencies are present [33], but not well documented during the annotation process due to resource constraints. To address these limitations, we introduce an *elaboration* component into our argument relation classification setting. This innovative approach leverages the generative capabilities of large language models to not only label the relationship between arguments but also to automatically produce concise, textual elaborations on why particular relations are assigned to claim and premise arguments. More specifically, with these elaborations, we aim to dissect the underlying argumentative reasoning of each argument pair, offering insights into the synthetic decision-making process and emulation of human annotation efforts to enhance the transparency of argument mining systems. Figure 2 illustrates a simplified elaboration example.

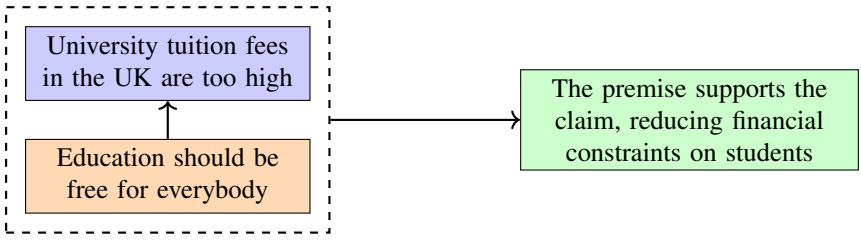

Figure 2. Example elaboration (green) about presented claim (blue) and premise (orange).

This methodology serves a general mitigation strategy of the opacity inherent to standard sequence classification models, allowing experts to understand and trust a model's final verdict by providing a parallel layer of observability. For elaborations to be effective, we utilize the generative capabilities of state-of-the-art LLMs as a function *ela* that is only dependent on the classification task description and the label scheme. An advantage of increasingly large pre-trained LLMs is the option to rely solely on few-shot inference settings and avoid fine-tuning the weights of the generative model if computational limitations permit it. In constrained scenarios, it would be feasible to generate elaborations based on a subset of the original data with an LLM and fine-tune a significantly smaller LLM based on the synthetic outputs in a knowledge distillation setting [34]. However, it is important to note that LLMs are inherently stochastic, which means the generated elaborations might vary even when the same input is provided, potentially introducing inconsistencies that could complicate transparency efforts. Nonetheless, generative outputs can be made more deterministic by configuring sampling parameters such as *temperature*, top-k, and top-p.

To address the challenge of evaluating the quality of elaborations, we train two parallel classification models: an argument-based model $model_a$ and an elaboration-based model $model_e$, both with the same training objective. The argument-based model receives the claim c and premise p, while the elaboration-based model receives the elaboration on

the argument pair $ela(c, p)$. This setup allows us to indirectly assess the quality of over one million elaborations without immediate human annotation, which can be introduced later on a selected subset of the elaborations.

4. Experimental Setup

This section targets the experimental assessment of the argument relation classification detection capabilities of benchmark models using the previously introduced KIALOPRIME dataset. Our setup describes two models: the argument sequence classification model $model_a$ based on the relations in the dataset and an elaboration sequence classification model $model_e$ that classifies synthetic elaborations on the argumentative interaction. Both models are evaluated under fixed conditions to ensure a fair performance comparison. Additionally, we evaluate how well each classification model can generalize during training over two variations.

4.1. Sequence Classification

The sequence classification models for both $model_a$ (using arguments as input) and $model_e$ (using elaborations as input) utilize the DeBERTa base architecture. In the argument sequence classification model, claim-premise pairs from the KIALOPRIME dataset are formatted as single sequences, with each pair separated by a special token. Both models are optimized with a learning rate of $3e-5$, using the AdamW optimizer, cross-entropy loss function, and a batch size of 100. They are trained for 5 epochs on the training split and evaluated on the validation split after each epoch. The model with the lowest validation loss is then selected for evaluation on the test split.

4.2. Elaboration Generation

Parallel to the argument sequence classification task, the elaboration sequence classification model is fed synthetic elaborations based on the argument pairs. The elaborations are generated by Mixtral 8x7B [35], an open-source mixture of experts LLM, served with vLLM [36] for efficient inference. To make the outputs more deterministic, we set the temperature to 0, enabling greedy sampling, which prioritizes the token with the highest probability. Additionally, we used a fixed random seed of 239 to ensure consistency across runs. The top-k parameter was set to -1 and top-p was set to 1, ensuring that all available tokens are considered without restricting the token pool. The generative model assesses the relationship between claims and premises while providing insights through elaborations on the argumentative interactions. This process operates in a few-shot setting to avoid extensive fine-tuning on our domain-specific dataset. The prompt includes a task description, static examples for each label (few-shot), and elaboration instructions. Generating these elaborations is computationally intensive, requiring four H100 GPUs and taking 12 hours to process over one million data points from the KIALOPRIME dataset.

4.3. Robustness

To assess the robustness of our models, specifically their ability to handle new contexts, we employ two distinct training split variations within the KIALOPRIME dataset. The first approach, referred to as *cross* involves distributing arguments of each discussion graph across the training, validation, and test sets. This method aims to ensure that all aspects of the dataset are represented in each phase of model training, providing a comprehensive exposure to the data's variability. In contrast, the second method, named *within* allocates entire discussion graphs to one of the training, validation, or test sets. This strategy is designed to minimize the leakage of discussion context information between splits and to more accurately gauge the dataset's representativeness and the models' ability to generalize. This approach tests the models' effectiveness in recognizing and adapting to entirely new contexts, as opposed to merely recalling specific discussion elements. The presented variations do not leak information about a specific pair of arguments and their relation.

Models are rigorously evaluated using a suite of standard performance metrics, which include cross-entropy loss (L), accuracy (A), F1-score macro (F1), precision (P), recall (R), and the Matthews correlation coefficient (MCC). These metrics allow us to quantitatively measure the models' performance and their ability to generalize across dataset variations.

5. Results and Discussion

Our experimental design evaluates the effectiveness of two distinct models for classifying argument relationships within the KIALOPRIME dataset: a traditional sequence classification model based on the directly presented arguments and an alternative sequence classification model based on the presented innovative elaboration component. In the following, we discuss the performance of these models and the implications of incorporating the elaboration component in our approach.

5.1. Text Statistics

Table 4 provides an initial comparison between the data based on the original arguments and the generated elaboration, which might provide hints about how much a model can learn from the presented information. Due to computational constraints, we generated elaborations with the cue to write as few sentences as necessary for describing the argumentative interaction. We observe that the distribution of token length averages is lower for the elaboration split, suggesting they are generally more concise, focusing on explaining the central points. The higher minimum and maximum values suggest that there is a required minimal length to elaborate on the interaction of two arguments, even if these two arguments are only of limited length. The total token counts highlight a substantial text volume in each split, with arguments having a significantly larger corpus. This analysis reveals the structural differences between argumentative and explanatory texts regarding token distribution.

Table 4. Text statistic comparison based on DeBERTa tokenization

Statistic	Arguments	Elaboration
Mean Length	80.39	55.82
Minimum Length	3	10
Maximum Length	774	998
Median Length	62.0	53.0
Total Count	97,724,916	67,858,964

5.2. Classification Results

Table 5 presents the classification results of both modeling approaches. They indicate robust performance in both training splits, with a slight advantage in the cross-discussion setting. The argument sequence classification model achieved an F1 of 0.908 and an MCC of 0.864 in the cross-discussion setup, compared to 0.899 F1 and an MCC of 0.851 within discussions. These metrics suggest that the model is capable of generalizing well within and across the discussions in the dataset. The elaboration sequence classification model displays a similar pattern of results. This model achieved an F1 of 0.843 and an MCC of 0.765 within discussions and 0.841 F1 with an MCC of 0.764 in cross-discussion evaluations. The slightly lower performance of the elaboration model may be attributed towards our approach of making the generated elaborations more deterministic by setting the sampling parameters, such as a temperature, to 0. While these settings ensure reproducibility and consistency, they also limit the model's creativity and ability to explore diverse reasoning paths. Introducing more randomness into the generation process might enhance the model's ability to produce richer, more varied elaborations, potentially improving its performance. However, this comes at the cost of making the results less reproducible, more complicated to analyze, and even more prone to hallucination.

Table 5. Results of Argument Relation Classification

Classification	Context	L	A	F1	P	R	MCC
Argument	Within	0.407	0.901	0.899	0.900	0.901	0.851
	Cross	0.398	0.910	0.908	0.910	0.910	0.864
Elaboration	Within	0.616	0.843	0.840	0.841	0.843	0.765
	Cross	0.658	0.842	0.841	0.844	0.842	0.764

5.3. Analysis of Elaboration Impact

We utilize generative reasoning to provide transparency for each relation classification decision, aiming closely at how human reasoning over argumentative structures. This approach targets the enhancement of understanding and trust by enabling users to analyze algorithmic decisions through textual elaborations. As the results show a slight decrease in quantitative performance metrics, we performed an explorative case analysis on the elaborations. Out of the generated elaborations, we selected a nuanced example in Figure 3 to highlight a common issue. In this instance, the model identified a neutral relationship between the claim and the premise. However, the participant is likely attacking a

claim that is not being made here, by equating meat with higher protein intake and arguing for the health benefits of a high-protein diet. This overlooks the fact that the original claim specifically targets the health effects of red meat, not protein-based diets in general. The generative model detects that the user is not addressing the original claim and thus assigns a neutral relation. This example underscores a key area for improvement: the model's ability to recognize flawed arguments and elaborate on implicit argumentative aspects. Enhancing the model to better capture these subtleties, potentially through improved prompt engineering and fine-tuning, could significantly improve its accuracy and reliability. By addressing these issues, the model can provide more contextually aware and interpretable decisions and potentially improve the results.

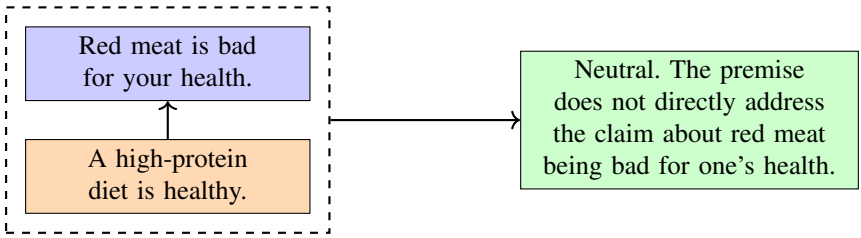

Figure 3. Selected elaboration assesses a neutral relation, while the ground truth is attack.

5.4. Implications for Argument Mining

The introduction of an elaboration component significantly advances the interpretability of argument mining systems, setting a new standard for developing complex systems where understanding the rationale behind decisions is crucial. Our approach proposes a dual model framework, with one model for the target objective and another one for elaborating, to enhance interpretability across various domains. In our experiments, we evaluated the utility of synthetic elaborations in argumentative interactions, identifying areas for improvement such as detecting potentially flawed argumentation and enhancing elaboration quality. The results suggest that synthetic elaborations can offer deeper insights into argumentative structures, making them valuable as annotation assistance. Integrating generative models enhances both the practical utility and interpretability of argument mining systems, ensuring accuracy and transparency. In the legal domain, for instance, this integration can improve the analysis and interpretation of legal texts by automatically generating explanations for model decisions. Our research demonstrates the benefits of generative models, contributing to the development of more sophisticated and interpretable argument mining systems for complex domains.

6. Conclusion and Future Work

In this study, we introduced KIALOPRIME, a novel large-scale benchmark dataset derived from the structured debates of the Kialo discussion platform. We applied topic modeling to identify a diverse range of topics within the dataset and analyze its characteristics. We formulated a strategy for constructing neutral relations to enable a tertiary classification problem setting. We demonstrated the utility of the presented dataset by explor-

ing the concept of synthetic elaborations on argumentative interaction designed to enhance the transparency of traditionally more opaque argument mining systems. To generate these elaborations, we employed the state-of-the-art Mixtral 7x8B mixture of experts LLM. We implemented and evaluated two benchmark tasks on our dataset: sequence classification on the arguments themselves and sequence classification on the generated elaborations. We evaluated the robustness of models trained on this dataset both across and within discussion graphs, ensuring no overlap of relations between the train, test, and validation splits. Our findings indicate minimal performance degradation between these two training setups for both benchmark models, reinforcing the dataset's value in capturing a wide range of discussions across various domains and topics. This robustness underscores KIALOPRIME's potential as a versatile tool for research and application in diverse argumentative environments. We identify areas of improvement, such as handling potentially flawed argumentation in online discussions when generating elaborations. Motivated by these findings, we aim to quantify argument quality and integrate this with generative models into end-to-end argument mining systems. Additionally, there are many references to external web pages available in the arguments, which we omitted in the scope of this study. Future work may utilize these references and apply retrieval-augmented generation to provide more contextualized and informed elaborations when observing relevant background information. Additionally, we plan to evaluate synthetic elaborations from a more theoretically-grounded reasoning perspective such as argument stances [37] and logical mechanisms [38], which we expect to improve the modeling results significantly. Expanding the use of LLMs to encompass both classification and generative tasks within this dataset will also be a key focus on our argument mining efforts.

Acknowledgments This research is funded by the Federal Ministry of Education and Research (BMBF, reference: 03RU2U151C) in the scope of the research project newspolygraph.

References

[1] Lawrence J, Reed C. Argument mining: A survey. Computational Linguistics. 2020;45(4):765-818.
[2] Dusmanu M, Cabrio E, Villata S. Argument mining on Twitter: Arguments, facts and sources. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017. p. 2317-22.
[3] Chakrabarty T, Hidey C, Muresan S, Mckeown K, Hwang A. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. In: EMNLP-IJCNLP 2019; 2019. p. 2933-43.
[4] Al Khatib K, Wachsmuth H, Kiesel J, Hagen M, Stein B. A news editorial corpus for mining argumentation strategies. In: Proceedings of COLING 2016; 2016. p. 3433-43.
[5] Ruiz-Dolz R, Nofre M, Taulé M, Heras S, García-Fornes A. Vivesdebate: A new annotated multilingual corpus of argumentation in a debate tournament. Applied Sciences. 2021;11(15):7160.
[6] Hautli-Janisz A, Kikteva Z, Siskou W, Gorska K, Becker R, Reed C. Qt30: A corpus of argument and conflict in broadcast debate. In: LREC; 2022. .
[7] Alhamzeh A, Bouhaouel M, Egyed-Zsigmond E, Mitrović J, Brunie L, Kosch H. A Stacking Approach for Cross-Domain Argument Identification. In: DEXA. Springer; 2021. .
[8] Lawrence J. Explainable argument mining. University of Dundee; 2021.
[9] Stab C, Gurevych I. Parsing argumentation structures in persuasive essays. COLING 2017.
[10] Cabrio E, Villata S. Five years of argument mining: A data-driven analysis. In: IJCAI; 2018. .
[11] Morio G, Ozaki H, Morishita T, Yanai K. End-to-end argument mining with cross-corpora multi-task learning. Transactions of the Association for Computational Linguistics. 2022;10:639-58.

[12] Ajjour Y, Chen WF, Kiesel J, Wachsmuth H, Stein B. Unit segmentation of argumentative texts. In: Proceedings of the 4th Workshop on Argument Mining; 2017. p. 118-28.
[13] Haddadan S, Cabrio E, Villata S. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics; 2019. .
[14] Bao J, He Y, Sun Y, Liang B, Du J, Qin B, et al. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In: Proceedings of the EMNLP 2022 Conference on Empirical Methods in Natural Language Processing;. .
[15] Gemechu D, Reed C. Decompositional argument mining: A general purpose approach for argument graph construction. In: Proceedings of the 57th Annual Meeting of the ACL; 2019. p. 516-26.
[16] Lenz M, Sahitaj P, Kallenberg S, Coors C, Dumani L, Schenkel R, et al.. Towards an Argument Mining Pipeline Transforming Texts to Argument Graphs;.
[17] Fromm M, Faerman E, Seidl T. TACAM: topic and context aware argument mining. In: IEEE/WIC/ACM International Conference on Web Intelligence; 2019. p. 99-106.
[18] Stab C, Miller T, Schiller B, Rai P, Gurevych I. Cross-Topic Argument Mining from Heterogeneous Sources. In: Proceedings of the EMNLP 2018;. .
[19] Agarwal V, Joglekar S, Young AP, Sastry N. GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. In: Proceedings of the ACM Web Conference 2022;. .
[20] Stab C, Gurevych I. Annotating Argument Components and Relations in Persuasive Essays. In: Tsujii J, Hajic J, editors. Proceedings of COLING 2014;. .
[21] Lawrence J, Visser J, Reed C. An online annotation assistant for argument schemes. In: Proceedings of the 13th linguistic annotation workshop. Association for Computational Linguistics; 2019. p. 100-7.
[22] Gorska K, Siskou W, Reed C. Annotating very large arguments. In: 9th International Conference on Computational Models of Argument 2022. IOS Press BV; 2022. p. 357-8.
[23] McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82.
[24] Duthie R, Lawrence J, Budzynska K, Reed C. The CASS technique for evaluating the performance of argument mining. In: Proceedings of the Third Workshop on Argument Mining (ArgMining2016);. .
[25] Ruiz-Dolz R, Taverner J, Lawrence J, Reed C. NLAS-multi: A Multilingual Corpus of Automatically Generated Natural Language Argumentation Schemes. arXiv preprint arXiv:240214458. 2024.
[26] Dhole KD, Agichtein E. Genqrensemble: Zero-shot llm ensemble prompting for generative query reformulation. In: European Conference on Information Retrieval. Springer; 2024. p. 326-35.
[27] Peng B, Quesnelle J, Fan H, Shippole E. YaRN: Efficient Context Window Extension of Large Language Models;. Available from: http://arxiv.org/abs/2309.00071.
[28] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al.. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks;. Available from: http://arxiv.org/abs/2005.11401.
[29] Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection;. Available from: http://arxiv.org/abs/2310.11511.
[30] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology. 2023.
[31] Grootendorst M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure;.
[32] Ruiz-Dolz R, Alemany J, Barberá SMH, García-Fornes A. Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation;36(6):62-70.
[33] Miura R, Tochigi A, Itoh T. Observation and Visualization of Subjectivity-based Annotation Tasks. In: 2022 26th International Conference Information Visualisation (IV). IEEE;. p. 85-90.
[34] Xu X, Li M, Tao C, Shen T, Cheng R, Li J, et al.. A Survey on Knowledge Distillation of Large Language Models;.
[35] Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al.. Mixtral of Experts;.
[36] Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu CH, et al.. Efficient Memory Management for Large Language Model Serving with PagedAttention;.
[37] Macagno F, Walton D, Reed C. Argumentation Schemes;. p. 517-74.
[38] Jo Y, Bang S, Reed C, Hovy E. Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes;9:721-39.