TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

A Comparative Study of debate speech analysis using Argumentation Mining in LLMs

Shweta Agarwal[a], Megha Agarwal[b], and Shobhit Sinha[c]

[a,b,c] Shri Ramswaroop Memorial University, Barabanki, India, 225003

[a]khetan.shweta@gmail.com, [b]meghaaggarwal.csis@srmu.ac.in, [c]sinha.shobhit@gmail.com

KEYWORD

Relation-based Argument Mining; Argument Mining; Large Language Model; Machine Learning; BERT

Abstract

Argument mining (AM) is an automated course of mining arguments, their mechanisms, and relationships from textual data. As online debate platforms become increasingly prevalent, there is a growing demand for effective AM techniques, particularly to support subsequent analytical tasks. One specific area within AM is Relation-Based Argument Mining (RbAM), which aims to identify agreement (support) and disagreement (attack) relationships among arguments.

However, RbAM presents significant challenges, and existing methods often fall short of achieving satisfactory performance in this domain. Our experiments focus on two open-source LLMs, namely Llama-2 and Mistral, which consist of ten files each. By harnessing the capabilities of these LLMs, we aim to address the limitations of current RbAM methods and pave the way for more effective and efficient argument mining techniques. Our findings underscore the potential of LLMs in advancing the state-of-the-art in RbAM and contribute to the broader goal of enhancing argument understanding and analysis in the era of online debates and discussions.

1. Introduction

Argument mining (AM) is the procedure of automatically mining arguments, their workings, and relationships from natural language text, as defined by scholars. The overarching AM problem encompasses three primary tasks:

a. **Argument Identification:** This task involves segmenting textual data into units and discerning which of these units are argumentative in nature.
b. **Identifying components of an Argument:** This task stereotypically revolves around categorizing claims and/or grounds within contrary text, distinguishing between different components of arguments.

Corresponding Author: Shweta Agarwal, Department of CSIS, Shri Ramswaroop Memorial University, Barabanki, India
Email: khetan.shweta@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

c. **Identification of Argumentative Relations:** Here, the goal is to determine how dissimilar writings are interconnected in a contrary discourse, specifically focusing on the relationships between arguments.

With the proliferation of platforms supporting online debates, there is an escalating demand for effective AM techniques, as noted by Lawrence and Reed (2019).

This paper zeroes in on a specific subset of AM falling under the third group, which aligns with the structure of debates found on platforms like kialo.com. On these platforms, arguments, represented as written remarks, are linked via support or attack relations.

More precisely, our research centres on relation-based Argument Mining (RbAM), a subfield where the primary task involves determining the relationship between given pairs of texts. In this context, given a pair (A, B) of texts A and B, the objective is to ascertain whether A attacks or supports B. This line of inquiry has been explored by various researchers, shedding light on the complexities and challenges inherent in RbAM tasks.

Consider the following three opinions extracted from the Debatepedia/ Procon dataset:
a1='Aborting a baby must be permitted',
a2='A baby must not be born until it is wanted', and
a3='Abortion leads to developing breast cancer in women'.

In this scenario, argument a2 can be construed as supporting argument a1, as it provides a rationale or justification for the assertion made in a1. Conversely, argument a3 can be viewed as attacking a1, as it presents a counterargument or challenges the assertion made in a1.

Relation-Based Argument Mining (RbAM) serves as a foundational component that underpins various downstream tasks within the realm of argumentation analysis. For instance, RbAM facilitates the process of gathering evidence to support or refute particular claims (Carstens and Toni, 2015). Additionally, it plays a crucial role in determining the acceptability and credibility of online arguments, a task vital for discerning the reliability and trustworthiness of information in digital discourse (Bosc et al., 2016). Moreover, RbAM is instrumental in analyzing contentious issues surrounding new regulations, enabling policymakers and stakeholders to better understand and navigate the complexities of public opinion and debate on regulatory matters (Konat et al., 2016). By providing insights into the relationships between arguments, RbAM empowers researchers and practitioners to delve deeper into the nuances of argumentative discourse and its implications across various domains.

Despite significant advancements in utilizing BERT-based models for Relation-Based Argument Mining (RbAM), it remains an interesting task. While some BERT-based models demonstrate promising performance on certain datasets, there is a lack of consistency in their effectiveness across different datasets. Studies have highlighted the same issue.

Cocarascu et al. (2020) found that while dissimilar BERT-based models performed rationally fine on specific datasets, separate baselines often struggled to deliver consistent results across multiple datasets. This inconsistency suggests that the effectiveness of BERT-based models for RbAM may be influenced by various factors, including dataset characteristics, domain-specific nuances, and the complexity of contrary discourse.

Similarly, Ruiz-Dolz et al. (2021) observed variations in the performance of BERT-based models across different datasets, indicating the challenges inherent in achieving robust and reliable performance in RbAM tasks using these models alone.

These findings underscore the need for further research and development efforts to address the challenges posed by RbAM. Future studies may explore novel approaches, including ensemble methods, domain adaptation techniques, and model fine-tuning strategies, to enhance the performance and generalizability of BERT-based models in tackling RbAM tasks effectively across diverse datasets and contexts.

This paper aims to leverage general-purpose Large Language Models (LLMs), suitably well-informed and encouraged, to tackle the Relation-Based Argument Mining (RbAM) task consistently across multiple datasets.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

Building upon recent studies that have showcased the superior performance of LLMs compared to existing baselines in other Argument Mining (AM) tasks, our helps can be outlined as follows:

A. **Proposing an Effective Methodology:** We introduce a novel methodology for conducting RbAM utilizing chat-based LLMs. By appropriately priming and prompting these LLMs, we aim to enhance their capability to understand and analyze argumentative discourse.
B. **Empirical Demonstration of Superior Performance:** Through extensive empirical evaluations conducted across ten datasets sourced from existing literature, we demonstrate the efficacy of our LLM-based approach for RbAM. Our method outperforms the current hi-tech RoBERTa baseline for RbAM.

By presenting these contributions, our paper seeks to advance the field of Argument Mining by harnessing the capabilities of LLMs and offering a robust and effective solution for addressing the challenges posed by RbAM tasks across diverse datasets and contexts.

2. Related Work:

In recent years, Relation-Based Argument Mining (RbAM) has emerged as a significant area of research, drawing substantial interest from scholars and practitioners alike (Cabrio and Villata, 2018). Various approaches have been proposed to address RbAM tasks, each with its strengths and limitations.

Hou and Jochim (2017) presented a Joint Inference model, which they associated in contrast to several baseline approaches, including logistic regression, attention-based LSTMs, and the EDITS method. Their approach outperformed these baselines, achieving an impressive F1 score of 65 on the Debatepedia/Procon dataset, which is too utilized in our study, albeit excluding the Procon arguments.

Cocarascu and Toni (2017) employed a deep learning planning featuring two distinct LSTMs on the embeddings of argument pairs, which were then concatenated using a softmax layer. Their method excelled on the Web-Content dataset, achieving an impressive F1 score of 89, a dataset that we also employ in our research.

Building on this, Cocarascu et al. (2020) explored 4 deep learning plannings with various embeddings, comparing them contrary to Random Forests and SVMs baselines. Despite achieving a macro F1 score of 54 across ten datasets, which are mostly utilized in our study, their performance was comparable to the baselines.

Trautmann et al. (2020) experimented with several LSTM variants, CAM-Bert, and TACAM-BERT on the UKP corpus (Stab et al., 2018), achieving their best F1 score of 80 with TACAM-BERT.

Jo et al. (2021) employed Logical Mechanisms and Argumentation Schemes, utilizing various models including TGA Net, Hybrid Net, BERT, BERT+Latent Cross, and BERT+Multi-task Learning as baselines. Their top-performing model achieved an F1 score of 77 on a dataset collected from the online debate site Kialo, also utilized in our study, and an F1 score of 80 on a dataset similar to Debatepedia/Procon (Cabrio and Villata, 2014).

Ruiz-Dolz et al. (2021) conducted an evaluation of numerous BERT-based models in contrast to LSTMs, accomplishing an F1 cut of 70 with RoBERTa-large on the US2016 debate corpus and the Moral Maze multi-domain corpus, mutually from AIFdb (non utilized here).

However, despite the progress made by these approaches, none of them leverage Large Language Models (LLMs) nor achieve satisfactory performance across datasets, which is the focus of our study. This underscores the need for further exploration and innovation in utilizing LLMs for Relation-Based Argument Mining tasks, an aspect we aim to address in our research.

2.1 Argument Extraction Using Large Language Models:

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

The extraordinary functioning of Large Language Models (LLMs) across a spectrum of Natural Language Processing (NLP) responsibilities has spurred inquiries into their efficacy in Argument Mining (AM) tasks. Thorburn and Kruger (2022) conducted a study where they fine-tuned GPT Neo, a pre-trained LLM, to produce natural language opinions either supporting or attacking a given matter dispute, demonstrating the potential of LLMs in this domain through prompt-based generation.

However, despite these advancements, Hinton and Wagemans (2023) noted that further research is necessary before LLMs can be deemed proficient in argumentative reasoning. This sentiment is echoed by Ruiz-Dolz and Lawrence (2023), who encountered challenges in their attempt to employ LLMs for detecting argumentative fallacies, with the models failing to outperform a RoBERTa-based Transformer model.

In the legal domain, Al Zubaer et al. (2023) fixated on the organization of dispute elements using GPT-3.5 and GPT-4 models, but found that LLMs could not exceed the performance of a area-explicit BERT-based baseline. However, there are hopeful developments, as demonstrated by Furman et al. (2023), who utilized LLMs to generate counter-narratives to combat online hate speech, achieving more favorable outcomes when complemented with argumentative strategies and analysis.

Overall, while LLMs show potential in AM tasks, there remain challenges and limitations that necessitate further research and refinement to fully harness their capabilities in reasoning and analyzing arguments effectively across various domains.

Van der Meer et al. (2022) showcased the potential of Large Language Models (LLMs) in Argument Mining (AM) tasks, specifically in the prediction of argument quality. Their study employed LLMs for tasks involving determining the validity and novelty of arguments. Notably, they achieved optimal performance by adopting a few-shot discovering ground plan with LLMs for the validity task, while a Transformer-based model fine-tuned for the innovation assignment yielded the finest results.

What distinguishes their work is the innovative approach of leveraging LLMs for AM tasks, which had not been extensively explored until then. Despite the substantial progress made in utilizing LLMs for various NLP tasks, the specific application of these models for Relation-Based Argument Mining (RbAM) had not been addressed in previous studies.

This observation underscores a critical gap in the existing research landscape. While LLMs have demonstrated their efficacy in a range of AM tasks, including argument quality prediction, their potential for RbAM remains largely unexplored. This highlights an exciting avenue for future research, where the capabilities of LLMs could be further investigated and harnessed for tasks related to analyzing argument relations, offering new insights and advancements in the field of Argument Mining.

3. Experimental Setup

We outline the datasets utilized, the baseline model for comparison, and the Large Language Models (LLMs) under experimentation.

Datasets: We utilized ten existing datasets, as follows:

a. Persuasive essays (Essay): Annotated corpus of 402 persuasive essays (Stab and Gurevych, 2017).
b. Microtexts (Mic): The corpus comprises 112 short texts that delve into controversial issues, covering a wide range of topics that spark debate and disagreement. Within these texts, a total of 576 arguments have been identified and extracted, providing a rich dataset for exploring diverse perspectives and viewpoints on contentious subjects. Each short text is likely to encapsulate a unique aspect or facet of the broader controversial issue, contributing to the breadth and depth of the corpus. Through the analysis of these arguments, researchers can gain insights into the multifaceted nature of controversial topics and the various arguments put forth by individuals engaging with these issues. (Peldszus and Stede, 2015).
c. Nixon-Kennedy debate (NK): The corpus originates from the 1960 Nixon-Kennedy presidential campaign, a pivotal moment in American political history. It focuses on five distinct topics relevant to the campaign, likely encompassing issues, events, or debates that were central to the electoral discourse

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

at that time. Given the historical significance of the Nixon-Kennedy campaign, this corpus provides valuable insights into the political climate, public sentiments, and key themes of the era. By examining the language and arguments present in this corpus, researchers can gain a deeper understanding of the strategies, narratives, and rhetorical tactics employed by political candidates and their supporters during this landmark election. (Menini et al., 2018).

d. Debatepedia-Procon (DP): The corpus is derived after 2 prominent working debate platforms: Debatepedia and Procon. Debatepedia is a collaborative wiki-style platform where users engage in structured debates on a wide range of topics, presenting arguments and counterarguments supported by evidence and reasoning. Procon, on the other hand, is a website dedicated to presenting both sides of contentious issues, featuring user-generated comments and arguments that reflect diverse viewpoints.

   By extracting data from these platforms, the corpus encompasses a comprehensive collection of arguments and discussions on various controversial topics. This dataset provides researchers with valuable insights into the nuances of public opinion, the diversity of perspectives, and the strategies employed in online debates. Analyzing this corpus can shed light on prevailing attitudes, societal values, and the dynamics of argumentation in contemporary discourse across different online platforms. (Cabrio and Villata, 2014).

e. IBM-Debater (IBM): The dataset comprises fifty five debated matters sourced from the debate motions database hosted on the International Debate Education Association (IDEA) website. These topics are carefully curated to encompass a wide spectrum of contentious issues that are likely to elicit divergent opinions and heated discussions.

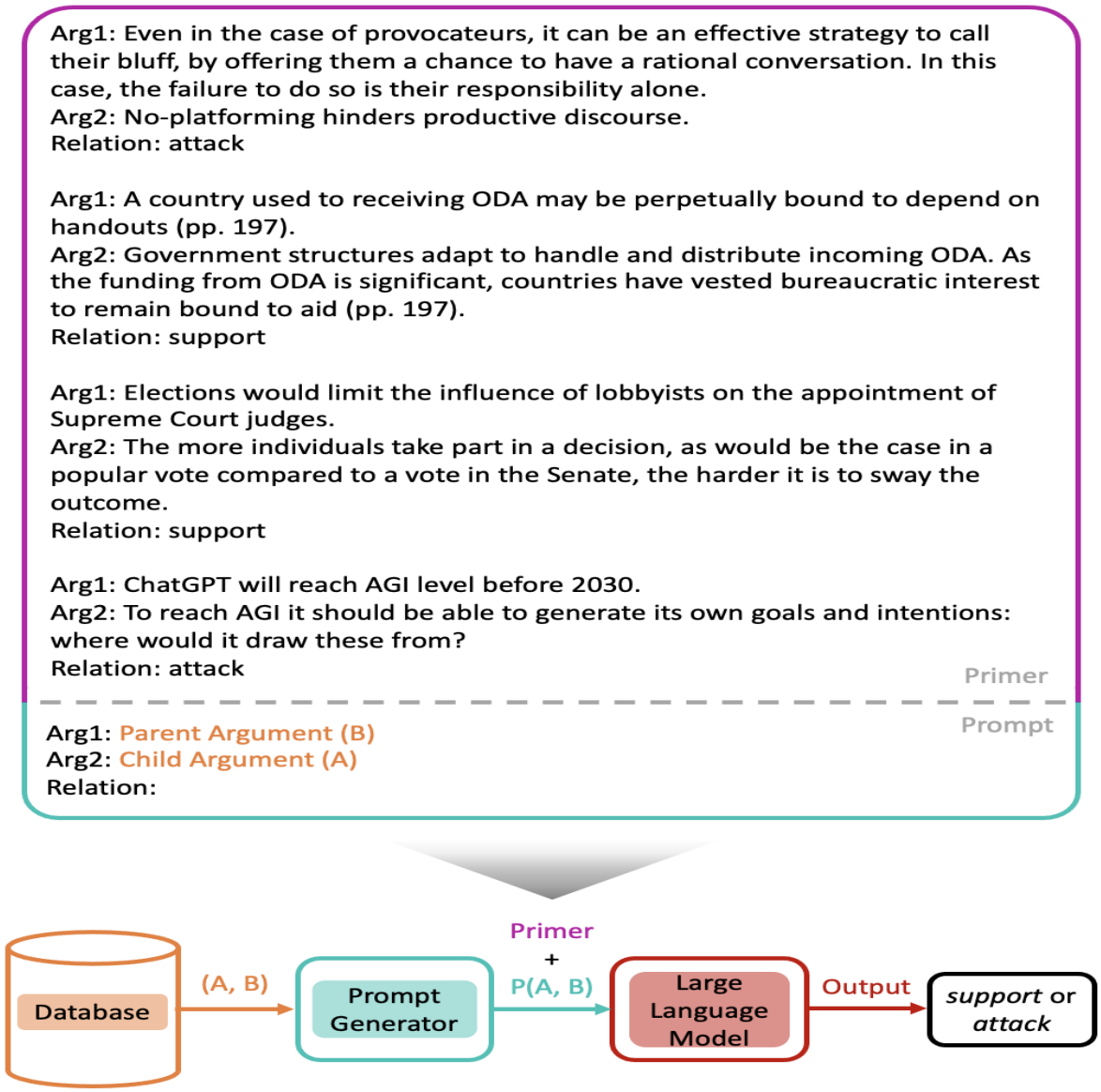
   IDEA is renowned for its commitment to promoting critical thinking, civil discourse, and debate education on a global scale. As such, the topics included in this dataset are likely to reflect a diverse range of societal, political, ethical, and cultural concerns that are relevant to contemporary discourse.

   By drawing from IDEA's extensive database, this dataset offers researchers a valuable resource for studying the dynamics of argumentation, examining public opinion, and exploring the complexities of controversial topics in various contexts. Analyzing this dataset can provide insights into the underlying reasoning, values, and perspectives that shape debates on important societal issues. (Bar-Haim et al., 2017).

f. ComArg: The corpus comprises user comments sourced from two platforms: Procon and IDEA. Procon, a consumer advocacy website, features a wide array of comments that provide insight into public opinions on various consumer issues. IDEA, an acronym for Intelligent Debate Enhancement Assistance, is a platform designed to facilitate and enhance structured online debates. The comments collected from these sources encompass diverse viewpoints and arguments, offering a rich dataset for analyzing user interactions and argumentative patterns in different contexts. By drawing from both Procon and IDEA, the corpus captures a broad spectrum of user-generated content, making it valuable for studying argumentation in consumer advocacy and online debate environments, adapted for the RbAM task (Boltužić and Šnajder, 2014).
g. CDCP: The corpus, which focuses exclusively on support relations, consists of 731 user comments concerning Consumer Debt Collection Practices. This dataset was meticulously annotated to identify and highlight supportive argumentative relationships within the comments. The annotations and structure of this corpus are based on the work of Park and Cardie (2018), who aimed to provide a detailed analysis of how users express support in discussions about debt collection practices. This specialized annotation allows for in-depth exploration of supportive arguments in consumer-related contexts.
h. UKP: Corpus with arguments obtained from web documents over eight controversial topics (Stab et al., 2018).
i. Web-Content (Web): This dataset comprises arguments that have been adapted from the Argument Corpus. In addition to these original arguments, the dataset includes a variety of arguments sourced from numerous other platforms, ensuring a diverse and comprehensive collection of argumentative content.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

j. Kialo: This dataset features debates extracted from Kialo, an online platform dedicated to structured debates. The arguments cover a wide range of topics, including Politics, Law, and Sports, providing a rich and varied set of perspectives and discussions.



Baseline: We selected RoBERTa, motivated by its strong performance in previous studies (Ruiz-Dolz et al., 2021). To fine-tune RoBERTa, we used 75% of each dataset, allocating this portion specifically for training. The fine-tuning process spanned 50 epochs, which allowed the model ample opportunity to learn the intricate patterns within the data. We set the batch size to 8, balancing computational efficiency and gradient stability. The learning rate was carefully chosen at 1e-5 to ensure that the model learned effectively without overshooting optimal parameter values. This fine-tuning strategy was designed to maximize RoBERTa's performance as a robust baseline for comparison with our selected LLMs. The best model (based on the highest F1 score on the validation set) for each dataset was selected for inference. The best model among all datasets (trained on Kialo) was chosen as the baseline.

3.1 Large Language Models:

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

I carefully selected two types of Large Language Models (LLMs) for our study, both of which are open-source:

- Llama 2 models: Pretrained with 2 trillion tokens, including Llama 2 13B and Llama 2 70B models.
- Mistral models: Including Mistral 7B and Mixtral 8x7B models. The Mixtral model combines eight Mistral 7B models.

All LLMs were experimented with both their base versions and GPTQ quantised versions, where individually weight is saved in 4 bits on the GPU, providing a trade-off between accuracy and space.

4. Results:

	RoBERTa	Llama13B	Llama13B-4bit	Llama70B-4bit	Mistral7B	Mixtral-8x7B-4bit
Essays	85 / 38 / 80	87 / 31 / 82	91 / 36 / 86	94 / 52 / **90**	89 / 42 / 85	94 / 43 / 89
Nixon-Kennedy	56 / 67 / 62	67 / 12 / 39	66 / 5 / 34	64 / 71 / **68**	54 / 68 / 61	66 / 50 / 58
CDCP	75 / - / 75	87 / - / 87	94 / - / **94**	92 / - / 92	75 / - / 75	93 / - / 93
UKP	68 / 81 / 75	70 / 82 / 77	75 / 84 / 80	84 / 89 / **87**	78 / 83 / 81	81 / 84 / 83
Debatepedia/Procon	90 / 89 / 90	83 / 71 / 77	84 / 72 / 79	96 / 95 / **96**	90 / 89 / 90	94 / 93 / 94
IBM-Debater	85 / 82 / 83	81 / 66 / 75	88 / 82 / 85	94 / 92 / 93	89 / 89 / 89	95 / 93 / **94**
ComArg	71 / 74 / 72	68 / 62 / 65	70 / 58 / 65	77 / 56 / 68	56 / 71 / 63	79 / 73 / **76**
Microtexts	73 / 53 / 67	76 / 45 / 67	84 / 41 / 72	81 / 52 / **73**	71 / 54 / 67	80 / 45 / 70
Web-Content	67 / 67 / 67	66 / 63 / 64	68 / 53 / 60	72 / 72 / **72**	57 / 72 / 64	70 / 66 / 68
Kialo	- / - / -	74 / 56 / 65	75 / 54 / 65	87 / 84 / **86**	83 / 83 / 83	85 / 82 / 84
Average	74 / 61 / 75	76 / 49 / 70	79 / 48 / 72	84 / 66 / **82**	74 / 65 / 76	84 / 63 / 81
Macro F_1	68	62	64	**75**	70	73
Inference Time (s)	0.005	0.11	0.34	1.73	0.06	0.28

Table 1: F_1 scores (as a percentage) for support / attack / both relations in various datasets (rows) for the models used (columns). RoBERTa here is the baseline (see §4) and boldface font indicates the best performing model (for both relations) for each dataset. The last row gives the time it takes for a single inference for each model, in seconds.

Table 1 showcases the comparative results of our study. Notably, the Llama 70B-4bit model attained the maximum macro F1 score of 75, outdoing all baseline models. Furthermore, this model secured the top F1 score in 7 from ten datasets (Essay, NK, UKP, DP, Mic, Web, and Kialo), surpassing all other LLMs and all but two baselines. Despite its impressive performance, the Llama 70B-4bit exhibited a relatively high inference time per disagreement pair is 1.73 seconds, likely due to its large size and the application of GPTQ quantization.

On the other hand, the Mixtral 8x7B-4bit model demonstrated nearly comparable performance using a macro F1 score of 73. While its typical F1 score for backup markers was on par with that of Llama 70B-4bit, it fell slightly behind in the average F1 score for attack labels. However, Mixtral 8x7B-4bit attained the maximum F1 value in the ComArg and IBM datasets. Additionally, it had a significantly faster inference period per argument pair is of 0.28 seconds, making it a more efficient option in terms of processing speed.

The Mistral 7B model, despite its smaller size, performed commendably, achieving a macro F1 score of 70, which surpassed all baseline models. Though, it did not outdo other large language models (LLMs) in any specific dataset. A notable advantage of Mistral 7B is its efficiency, with the fastest inference time per argument pair is of 0.06 seconds.

The Llama 13B-4bit and Llama 13B models attained alike macro F1 scores of 64 and 62, respectively, though their functioning varied across different datasets. Specifically, Llama 13B-4bit excelled in the CDCP dataset, likely due to its tendency to more frequently generate support labels. The performance of both models saw a slight improvement with the application of GPTQ quantization. Despite these enhancements, both models still underperformed compared to the best baseline models. An unexpected finding was that Llama 13B-4bit had a slower inference time compared to the non-quantized Llama 13B.

5. Conclusion and Future Work

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

We have developed a novel approach for tackling the Relation-Based Argument Mining (RbAM) task by leveraging versatile Large Language Models (LLMs), which are effectively briefed and stimulated for this purpose. In our extensive experimentation, conducted across ten different datasets and involving 5 open-source LLMs (with the majority being quantized), we found significant performance improvements over the traditional RoBERTa baseline. Specifically, the Llama 70B-4bit and Mixtral 8x7B-4bit models demonstrated superior performance, with Llama 70B-4bit emerging as the best-performing model. However, it is important to note that this model also demands more GPU resources and has slower inference times compared to the other models tested.

For future research, several potential avenues can be explored:

- Entity Masking: Highlighting the argumentative structure of sentences by masking entities has been demonstrated to improve performance in the argument retrieval task, as evidenced by the findings of Ein-Dor et al. (2020). This technique involves replacing specific entities within sentences with placeholders, which helps in emphasizing the underlying argumentative components, leading to better retrieval and analysis of arguments.
- Improvement of Attack Relation Prediction: Improving the accuracy of predictions regarding attack relations is essential, given that both large language models (LLMs) and traditional baseline methods have shown relatively poor performance in this area.
- Extension to Ternary RbAM Task: Extending this work to address difficult three- way RbAM task, which involves identifying if there exists a support, an attack, or no relation between two arguments.

By pursuing these avenues, we aim to further enhance the effectiveness and robustness of RbAM methods, contributing to advancements in argument mining and related fields.

6. Limitations

Our work has several limitations that warrant acknowledgment. Firstly, we focus on the binary Relation-Based Argument Mining (RbAM) task, which involves identifying support or attack relations. However, in many real-world scenarios, the task may involve a three-way RbAM task, which includes recognizing support, attack, or no relation between arguments. This limitation highlights the need for future research to address more diverse and complex argument mining tasks.

Additionally, our study is conducted primarily on English-language datasets. It remains uncertain whether Large Language Models (LLMs) will react equally good in RbAM tasks in different languages, underscoring the necessity for cross-linguistic evaluations and adaptations.

Furthermore, our choice of LLMs is influenced by GPU limitations, leading us to select smaller or quantised models. Unfortunately, this constraint prevented us from fine-tuning any of the LLMs due to computational feasibility concerns. This limitation may impact the generalizability of our findings and underscores the need for further exploration with more extensive computational resources.

References

[1] Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. "Performance analysis of large language models in the domain of legal argument mining." Frontiers in Artificial Intelligence, 6, 2023.

[2] Andreas Peldszus and Manfred Stede. "Joint prediction in MST-style discourse parsing for argumentation mining." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 938–948, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[3] Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. "A corpus of argument networks: Using graph properties to analyze divisive issues." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3899–3906, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[4] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

https://doi.org/10.5281/zenodo.13294365

[5] Christian Stab and Iryna Gurevych. "Parsing argumentation structures in persuasive essays." Computational Linguistics, 43(3):619–659, September 2017.

[6] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. "Cross-topic argument mining from heterogeneous sources." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3664–3674, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[7] Christian Stab and Iryna Gurevych. "Parsing argumentation structures in persuasive essays." Computational Linguistics, 43(3):619–659, September 2017.

[8] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. "Cross-topic argument mining from heterogeneous sources." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3664–3674, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[9] Damián Ariel Furman, Pablo Torres, José A. Rodríguez, Diego Letzen, Maria Vanina Martinez, and Laura Alonso Alemany. "High-quality argumentative information in low resources approaches improve counter-narrative generation." In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore

[10] Dietrich Trautmann, Michael Fromm, Volker Tresp, Thomas Seidl, and Hinrich Schütze. "Relational and Fine-Grained Argument Mining: The LMU Munich project ReMLAV within the DFG Priority Program RATIO 'Robust Argumentation Machines'." Datenbank-Spektrum, 20(2):99–105, July 2020.

[11] Ein-Dor, Liat, et al. "Corpus wide argument mining—a working solution." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05. 2020.

[12] Elena Cabrio and Serena Villata. "Combining textual entailment and argumentation theory for supporting online debates interactions." In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 208–212, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[13] Elena Cabrio and Serena Villata. "Node: A benchmark of natural language arguments." In Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014, volume 266 of Frontiers in Artificial Intelligence and Applications, pages 449–450. IOS Press, 2014.

[14] Elena Cabrio and Serena Villata. "Five Years of Argument Mining: a Data-driven Analysis." In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pages 5427–5433, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization.

[15] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. "GPTQ: accurate post-training quantization for generative pre-trained transformers," 2022.

[16] Filip Boltužić and Jan Šnajder. "Back up your stance: Recognizing arguments in online discussions." In Proceedings of the First Workshop on Argumentation Mining, pages 49–58, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[17] Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. "Exploring the potential of large language models in computational argumentation," 2023.

[18] Jiang, Albert Q., et al. "Mixtral of experts." arXiv preprint arXiv:2401.04088 (2024).

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

[20] Joonsuk Park and Claire Cardie. "A corpus of eRulemaking user comments for measuring evaluability of arguments." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[21] Lucas Carstens and Francesca Toni. "Towards relation based argumentation mining." In Proceedings of the 2nd Workshop on Argumentation Mining, pages 29–34, Denver, CO, June 2015. Association for Computational Linguistics.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.01 (2024)
https://www.tejasjournals.com/
https://doi.org/10.5281/zenodo.13294365

[22] Luke Thorburn and Ariel Kruger. "Optimizing language models for argumentative reasoning." In Proceedings of the 1st Workshop on Argumentation & Machine Learning co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff,Wales, September 13th, 2022, volume 3208 of CEUR Workshop Proceedings, pages 27–44. CEUR-WS.org, 2022.

[23] Marco Lippi and Paolo Torroni. "Argumentation mining: State of the art and emerging trends." ACM Trans. Internet Techn., 16(2):10:1–10:25, 2016.

[24] Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. "A corpus for research on deliberation and debate." In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 812–817, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[25] Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. "Will it blend? mixing training paradigms & prompting for argument quality prediction." In Proceedings of the 9th Workshop on Argument Mining, pages 95–103, Online and in Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics.

[26] Oana Cocarascu and Francesca Toni. "Identifying attack and support argumentative relations using deep learning." In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1374–1379, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[27] Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. "Dataset independent baselines for relation prediction in argument mining." In Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020, volume 326 of Frontiers in Artificial Intelligence and Applications, pages 45–52. IOS Press, 2020.

[28] Ramon Ruiz-Dolz and John Lawrence. "Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models." In Proceedings of the 10th Workshop on Argument Mining, pages 1–10, Singapore, 2023. Association for Computational Linguistics.

[29] Ramon Ruiz-Dolz, Jose Alemany, Stella M. Heras Barberá, and Ana García-Fornes. "Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation." IEEE Intelligent Systems, 36(6):62–70, November 2021. Conference Name: IEEE Intelligent Systems.

[30] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. "Stance classification of context-dependent claims." In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 251–261, Valencia, Spain, April 2017. Association for Computational Linguistics.

[31] Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. "Never retreat, never retract: Argumentation analysis for political speeches." pages 4889–4896, 2018.

[32] Tom Bosc, Elena Cabrio, and Serena Villata. "Tweeties squabbling: Positive and negative results in applying argument mining on social media." 287:21–32, 2016.