DISPUTool 3.0: Fallacy Detection and Repairing in Argumentative Political Debates

Pierpaolo Goffredo, Deborah Dore, Elena Cabrio, Serena Villata
Université Côte d'Azur, CNRS, INRIA, I3S, France
{firstname.surname}@univ-cotedazur.fr

Abstract

This paper introduces and evaluates a novel web-based application designed to identify and repair fallacious arguments in political debates. DISPUTool 3.0 offers a comprehensive tool for argumentation analysis of political debate, integrating state-of-the-art natural language processing techniques to mine and classify argument components and relations. DISPUTool 3.0 builds on the *ElecDeb60to20* dataset, covering US presidential debates from 1960 to 2020. In this paper, we introduce a novel task which is integrated as a new module in DISPUTool, i.e., the automatic detection and classification of fallacious arguments, and the automatic *repairing* of such misleading arguments. The goal is to show to the user a tool which not only identifies fallacies in political debates, but it also shows how the argument looks like once the veil of fallacy falls down. An extensive evaluation of the module is addressed employing both automated metrics and human assessments. With the inclusion of this module, DISPUTool 3.0 advances even more user critical thinking in front of the augmenting spread of such nefarious kind of content in political debates and beyond. The tool is publicly available here: https://3ia-demos.inria.fr/disputool/

1 Introduction

Argumentation is the process by which arguments are constructed and handled: this means that arguments are compared, evaluated in some respect and judged to establish whether any of them is warranted. Argument Mining (AM) (Cabrio and Villata, 2018; Lawrence and Reed, 2019) is the research field in artificial argumentation aiming at automatically processing natural language arguments and reasoning upon them. It aims at extracting natural language arguments and their relations from text, with the final goal of providing machine-processable structured data for computational models of argument. More precisely, AM deals with the identification of argumentative components (i.e., premise, claim) and the prediction of the relations holding between these components (i.e., attack, support) in text. To further improve the quality of arguments (Wachsmuth et al., 2024), AM involves the identification and classification of fallacious arguments (Oswald and Herman, 2020). These arguments are defined as invalid or wrong moves in argumentative discourse (van Eemeren, 2015). The resulting argumentation is therefore misleading.

Once detected, fallacious arguments can be corrected to transform them into valid, non-fallacious arguments. We call this task *repairing fallacious arguments*. In this task, fallacious arguments are refined into a new version that is *clearer, fairer, and free from manipulative techniques*. This helps the audience to get a better understanding of the content of the argument and the impact of its misleading components in the argument interpretation.

In this paper, we present a novel version of DISPUTool, i.e., DISPUTool 3.0, which aims to automatically analyse political debates from the argumentation point of view.

In addition to the previous version of the tool, where argument components and relations were identified and analysed on the political debates of the US presidential campaigns from 1960 to 2016 (Goffredo et al., 2023a), we introduce a novel module where *i)* fallacious arguments are automatically identified and classified in the political debate proposed by the user, and *ii)* a non-fallacious reformulation of the fallacious argument is proposed to the user. The fallacy identification and repairing module employs advanced AM techniques to detect, classify and repair the fallacies.

For the task of *Fallacy Detection and Classification*, we employ MultiFusion BERT (Goffredo et al., 2023b). This transformer-based architecture integrates various engineered features to simultaneously detect and classify the fallacious argument into six different categories, i.e., *Ad*

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 472–480

Hominem, *Appeal to Emotion*, *Appeal to Authority*, *Slippery Slope*, *False Cause*, and *Slogan*. Once detected, each fallacious argument is transformed into a non-fallacious one using a Large Language Model (LLM). Currently, DISPUTool leverages LLama 3 8B (Dubey et al., 2024). LLama has been trained using specific prompt techniques on the **FallacyFix**[1] dataset. The arguments generated with the model are evaluated using both automatic and human evaluation metrics.

To the best of our knowledge, DISPUTool is the only automatic tool that integrates the identification and classification of argument components, relations, and fallacies within a single application. This tool represents a significant improvement towards the computational support for political debate analysis, offering both to scholars in social sciences and to general public users an effective way to achieve a better understanding of the underlying complexities of argumentation in political debates.

2 DISPUTool 3.0 Main Functionalities

In this section, we present DISPUTool's main functionalities, with a focus on the new module for fallacy detection and repairing. Additionally, DISPUTool 3.0 has been improved so that each processing step can be executed using our publicly accessible REST API, promoting reusability.

2.1 Dataset

DISPUTool 3.0 enables comprehensive analysis of U.S. televised presidential debates from 1960 to 2020, extending the coverage of the previous version which considered the debates from 1960 to 2016 only (Goffredo et al., 2023b). This new version of the dataset includes 44 debates, expanding upon the 39 included in the previous release. The *ElecDeb60to20* dataset has been annotated with argument components (*claim*, *premise*), relations between components (*attack*, *support*), and argumentative fallacies. In particular, we consider the following fallacies: *Ad Hominem* (when the argument becomes an excessive attack on an arguer's position), *Appeal to Authority* (when the arguer mentions an authority who agreed with her claim either without providing relevant evidence, or by mentioning popular non-expert), *Appeal to Emotion* (when there is an unessential loading of emotional language), *False Cause* (when there is a misinterpretation of the correlation of two events for causation), *Slippery Slope* (when it suggests that an unlikely exaggerated outcome may follow an act), and *Slogans*. Table 1 reports on the dataset' statistics.

	Classes	Instances	Distribution
Argument Components	Claim	29624	53%
	Premise	26055	47%
	Total	*55679*	*100%*
Argument Relations	Attack	21687	85%
	Support	3835	15%
	Total	*25522*	*100%*
Fallacious Argument Components	Ad Hominem	341	12%
	Appeal to Emotion	1591	58%
	Appeal to Authority	433	16%
	False Cause	179	7%
	Slippery Slope	122	4%
	Slogans	78	3%
	Total	*2744*	*100%*

Table 1: Statistics on the different annotation layers of the ElecDeb60to20 dataset.

The training dataset has been built from the official website of the Commission on Presidential Debates (CPD) [2], ensuring access to verified and complete debate transcripts.

2.2 Argumentative Structure Analysis

DISPUTool allows also to explore the argumentative structure of each debate. From the home page, users can select a specific debate year (e.g., McCain vs. Obama 2008). The number of debates varies by election cycle, with some years featuring more debates than others (e.g., three debates in 2020, four in 2000). Upon selecting a debate, the tool highlights key argumentative elements in the manually annotated *ElecDeb60to20* dataset:

- *Argument Components*: the components put forward by each candidate. The tool labels them as either *claim* or *premise*;
- *Argument Relations*: building upon version 2.0, the tool now offers an improved identification and classification of the relations between argumentative components, categorising them as either *support* or *attack*;
- *Fallacious Arguments*: DISPUTool 3.0 highlights, differently from its previous version, fallacious arguments in the *ElecDeb60to20* dataset (Goffredo et al., 2023b), identifying

[1] https://github.com/pierpaologoffredo/repairing_fallacies

[2] www.debates.org

the boundaries of the fallacy and categorising it into one of the following six categories: *Ad Hominem*, *Appeal to Authority*, *Appeal to Emotion*, *False Cause*, *Slogan*, *Slippery Slope*.

2.3 Data Exploration

Users can explore the manually annotated debates through multiple data visualisations, enhancing their understanding of the content.

Named Entity Recognition. *Word clouds* provide an intuitive representation of key terms, with font sizes reflecting word frequency. *Sankey diagrams* and *Stacked Area* charts allow users to visualize the identified Named Entities (NEs), extracted using the Stanford Named Entity Recognizer [3]. Users can filter results based on various criteria, including the type of NE, the year of the debate, and the name of the candidate. This visualization makes explicit what are the NE (e.g., Fidel Castro, Iraq war) employed the most in the discourses of each of the candidates to the presidential elections.

Fallacies. The user can explore the different argumentative fallacies of the dataset using *Sankey diagrams*. The tool allows filtering based on the year of the debate, on the type of fallacy the user is interested in, and on the name of the candidate that stated the fallacy. This visualisation lets the user compare two candidates debating against each other in terms of the kind and quantity of fallacious arguments they put forward. This provides an overview of each debate in terms of propagandist and fallacious arguments put forward in there.

2.4 Interactive Analysis and Fallacy Repair

The novel *AM, Fallacy Detection & Unveiling* module of DISPUTool 3.0 provides users with an interactive tool to analyze a debate they propose. More precisely, users can either select their own political debate text or choose one from a short list of suggestions. This module enables testing the proposed models on two different tasks: *i)* the automatic *detection and classification of argument components, argument relations, and fallacies*, and *ii)* the *repairing of the identified fallacies* in the political debate text by the user.

This functionality allows users to assess the accuracy and effectiveness of DISPUTool 3.0's algorithms in real-world scenarios, facilitating a deeper understanding of both the tool's capabilities and the investigation through our tool of the complexity of political argumentation in the debates they are interested in. In the following, we describe in detail the two models for argumentation analysis and fallacy repairing.

Debate Analysis. DISPUTool 3.0 introduces an enhanced functionality, allowing users to automatically analyse political debate texts with higher precision. The tool automatically detects and classifies *argument components*, *argument relations*, and—new in this version—*fallacies*. Arguments are systematically labeled as either *premises* or *claims*, while fallacies are not only highlighted but also categorised based on their type. Additionally, the tool generates a visual graph that maps the relationships between arguments, differentiating between *supporting* and attacking *relations*. This graphical representation provides users with a clearer understanding of the argumentative structure within the debates.

Fallacy Repair. DISPUTool 3.0 introduces a novel and unique, to the best of our knowledge, functionality: the repairing of fallacious arguments. When the user provides as input a political debate text containing a fallacy, the system performs a two-step process:

1. *Fallacy Detection and Classification*: It analyses the text, highlighting the boundaries of the fallacious argument, and it determines its specific type among a provided set of six fallacy categories.

2. *Fallacious Argument Repairing*: Once the fallacy is identified, the system generates a revised version of the text, where the fallacious argument is replaced with a logically sound counterpart that aims at being *clearer, fairer, and free from any technique that could negatively persuade the audience*.

This module provides a practical demonstration of how flawed arguments can be restructured to improve their logical validity, offering a valuable learning tool for users to understand the nuances of sound argumentation, and promoting a healthier political discourse.

3 Evaluation and Results

The DISPUTool architecture is visualized in Figure 1. In this section, we describe the experimental setting for evaluating the performance of the

[3] https://nlp.stanford.edu/software/CRF-NER.html

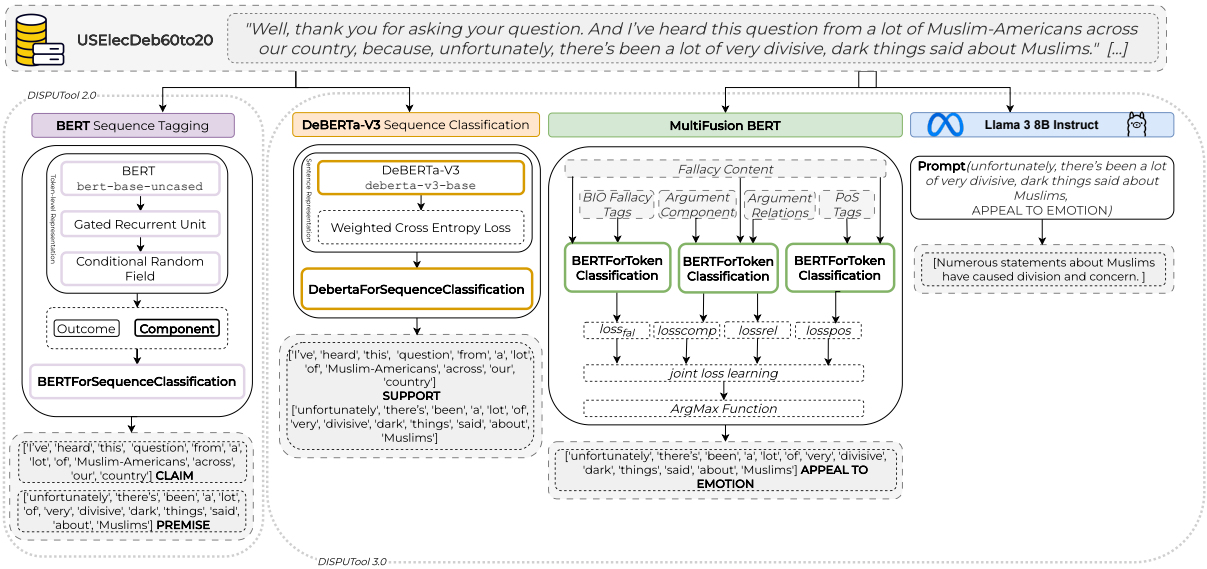

Figure 1: DISPUTool 3.0 new architecture.

argument component detection and classification, argument relation prediction, and fallacy detection and repairing modules.

Argument Detection and Classification. This section reports the results we obtained for the task of *Argument Component Detection and Classification* (Goffredo et al., 2023a). For this task, we followed the architecture proposed by Mayer et al. (2021), framing argument component detection as a sequence tagging task using the BIO tagging scheme. Sentence representations at the token level were computed with a fine-tuned BERT model (Devlin et al., 2019), trained for 15 epochs using the Adam optimizer, with a learning rate of 6e-5 and a maximum sequence length of 64. These representations were then fed into a Gated Recurrent Unit (GRU) (Cho et al., 2014), followed by a Conditional Random Field (CRF) (Lafferty et al., 2001). The dataset was split into training (80%), validation (10%), and test (10%) sets. The final model achieved an F1-score of 0.79 on the test set.

Relation Prediction. DISPUTool 3.0 leverages a fine-tuned DeBERTa-V3 (He et al., 2023) model to detect and classify relations between arguments, i.e., *support* and *attack*. This task is framed as a three-class classification problem, where the model assigns a label within *support*, *attack* and *no-relation* to each pair of arguments.

The fine-tuning process was conducted over 3 epochs, employing a learning rate of 4e-5, a batch size of 16, and a maximum sequence length of 255 sub-word tokens. To mitigate the impact of class imbalance during training, the model incorporates a weighted Cross Entropy Loss adjusted to the distribution of the three classes.

This optimized configuration achieved a Macro F1-score of 0.69 on the test set, representing a substantial improvement over the previous DISPUTool version, which relied on a RoBERTa-based model (Zhuang et al., 2021) and reached a Macro F1-score of 0.60 (Goffredo et al., 2023a). A comprehensive comparison of all evaluated models is provided in Table 2.

Model	Method	Macro F1 Score
DistilBERT	seq-class	0.581
BERT	sent-class	0.590
DISPUTool 2.0's RoBERTa	seq-class	0.601
XLM-RoBERTa	seq-class	0.637
BERT	seq-class	0.664
DeBERTa	seq-class	**0.690**

Table 2: Results of Relation Prediction task based on sequence classification among the labels {*Support, Attack, NoRel*}.

All tested models were hyperparameter-tuned using the Argumentation Mining Transformers Module (AMTM) [4] over the following ranges: number of epochs $\in 1, 2, 3$, batch size $\in (8, 16, 32)$, maximum sequence length $\in (128, 256, 512)$, and learning rate $\in (1e^{-5}, 2e^{-5}, 3e^{-5}, 4e^{-5})$.

Fallacy Detection and Classification. For the task of fallacy detection and classification, we em-

[4] https://github.com/crscardellino/argumentation-mining-transformers

ployed MultiFusion BERT (Goffredo et al., 2023b), a transformer-based architecture that integrates the text of the debate, its argumentative features (i.e., *components* and *relations*), and various engineered features for the task.

MultiFusion BERT[5] exploits three specialised *TokenForClassification* Transformer models to perform distinct tasks. One model is dedicated to detect and classify fallacies, another model handles argumentative features by processing both components and relations, and a third model focuses on the part-of-speech (PoS) tags.

The system computes separate losses for each task: $loss_{fal}$ for fallacy detection, $loss_{cmp}$ and $loss_{rel}$ for argument components and relations respectively, and $loss_{pos}$ for PoS tagging. These losses are then combined using a weight factor of $\alpha = 0.1$ into a unified joint loss. The model employs the Adam optimizer with gradient clipping at a maximum norm of 10, dropout of 0.1, a learning rate of 4×10^{-5}, and batch sizes of 8 for training and 4 for testing. Training involves four epochs for fine-tuning and optimisation.

Table 3 reports the evaluation results of MultiFusion BERT and other baseline models (see Goffredo et al. (2023b) for a comprehensive evaluation), highlighting their respective performance across the fallacy detection and classification task.

Model	**Macro F1 Score**
BERT + LSTM	0.469
BERT + LSTM *(comp. and rel. features)*	0.514
BERT + BiLSTM + LSTM	0.549
BERT + BiLSTM + LSTM *(comp. and rel. features)*	0.561
DistilbertFTC `distilbert-base-cased`	0.701
DistilbertFTC `distilbert-base-uncased`	0.704
BertFTC `bert-base-uncased`	0.709
DebertaFTC `microsoft/deberta-base`	0.722
MultiFusion BERT *(comp., rel. and PoS features)*	**0.739**

Table 3: Average macro F1 scores for fallacy detection (BIO labels are merged) using different models (FTC stands for "ForTokenClassification").

Repairing Fallacies. Prior research on fallacious argumentation has largely focused on detecting and classifying fallacies (Sahai et al., 2021; Alhindi et al., 2022; Goffredo et al., 2022, 2023b; Helwe et al., 2024; Chen et al., 2024; Alhindi et al., 2024). However, these approaches fall short of addressing the issue of *how to transform fallacious arguments into logically valid and fair statements*. To solve this problem, we introduce the task of *repairing fallacious arguments*, which aims to modify fallacious statements into versions that are *clearer*, *fairer*, and *free from manipulative techniques*.

To evaluate our proposed model on this task, we built a new resource, called **FallacyFix**, which comprises 747 repaired examples of fallacious arguments derived from the *ElecDeb60to20-fallacy* dataset (Goffredo et al., 2022). These repaired fallacious arguments span various fallacy categories, i.e., *Appeal to Fear*, *Appeal to Pity*, *Appeal to Popular Opinion*, *Flag Waving*, and *Loaded Language*. Table 4 presents different examples of repaired fallacies, and Table 5 shows the distribution of the different types of fallacies in the FallacyFix dataset.

Subcategory	**Frequency**	**Distribution**
Loaded Language	416	56%
Flag waving	147	20%
Appeal to Pity	83	11%
Appeal to Fear	61	8%
Appeal to Popular Opinion	40	5%
Total	*747*	*100%*

Table 5: Statistics of the FallacyFix dataset.

To address the repairing process, we put in place a systematic methodology, grounding on linguistics techniques such as *Population Reference Elimination*, *Emotional Content Subtraction*, and *Semantic Simplification*. These techniques are tailored to the linguistic features of each fallacy type to ensure that the repaired arguments keep their core meaning while eliminating the manipulative element(s).

To address the *repairing fallacies* task in an automatic way, we employed modular prompt-based techniques using LLMs such as GPT-4 (OpenAI, 2023) and Llama 3 8B (Dubey et al., 2024). These techniques were evaluated across three configurations: *i) Zero-Shot* (`ZS`) that relies on minimal input without examples; *ii) Few-Shot* (`FS`) that includes demonstrative examples, and *iii) Fine-Tuning* (`FT`) that incorporates task-specific training instructions. In the (`FS`) setting, examples are fixed thorough each trial.

We designed our prompt using modular components (see Figure 2), and we tested different configurations by including or excluding two fundamental elements: the *gold* fallacy label and the *contextual information* surrounding the fallacy requiring to be repaired (i.e., the arguments immediately before

[5] `https://huggingface.co/pierpaologo/MultiFusionBERT`.

Category	Context and Fallacious Argument	Repaired Argument	Strategy
Appeal to Pity	I think if you talk to anybody, it's not choice. *I've met people who struggled with this for years, people who were in a marriage because they were living a sort of convention, and they struggled with it.* And I've met wives who are supportive of their husbands or vice versa when they finally sort of broke out and allowed themselves to live who they were, who they felt God had made them.	I think if you talk to anybody, it's not choice. *I've met people who had issues for years.* And I've met wives who are supportive of their husbands or vice versa when they finally sort of broke out and allowed themselves to live who they were, who they felt God had made them.	Generalizing and weakening
Appeal To Popular Opinion	His has been in the legislative branch. *I would say that the people now have the opportunity to evaluate his as against mine and I think both he and I are going to abide by whatever the people decide.* Well, I'll just say that the question is of experience and the question also is uh - what our judgment is of the future, and what our goals are for the United States, and what ability we have to implement those goals.	His has been in the legislative branch. Well, I'll just say that the question is of experience and the question also is uh - what our judgment is of the future, and what our goals are for the United States, and what ability we have to implement those goals.	Removing additional reference
Flag Waving	And I only want to say that however good the record is, it's got to be better. *Because in this critical year - period of the sixties we've got to move forward, all Americans must move forward together, and we have to get the greatest cooperation possible between labor and management.* We cannot afford stoppages of massive effect on the economy when we're in the terrible competition we're in with the Soviets.	And I only want to say that however good the record is, it's got to be better. *Because in this critical year - it's necessary the greatest cooperation possible between labor and management.* We cannot afford stoppages of massive effect on the economy when we're in the terrible competition we're in with the Soviets.	Rephrasing
Loaded Language	I wasn't just getting more power and more power. *So I rolled the dice, I put my career on the line because I really believe the future of America is on the line.* We can give you all these numbers, they don't mean a thing.	I wasn't just getting more power and more power. We can give you all these numbers, they don't mean a thing.	Removing additional information

Table 4: Examples of fallacious arguments alongside their repaired versions and the strategies used for repair. Each fallacious argument is embedded in its context and shown in *italics*. The repaired version includes the same context, with the corrected argument also in *italics*.

and after the fallacious statement). The prompt

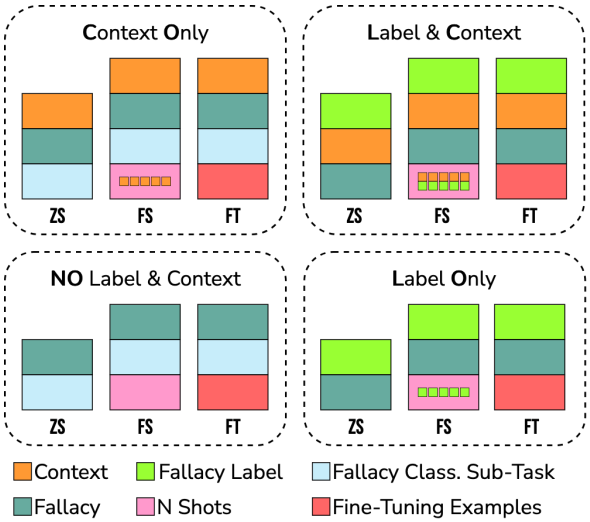

Figure 2: Prompt modularity based on the specific configurations and settings.

structure consists of a *context* section (when applicable) and a *core fallacious argument* section (i.e., the fallacy to be repaired). In total, we employed four distinct approaches to evaluate the model's performance in identifying and repairing fallacies:

- **Context Only (`CO`)**: we provided the model with the contextual information and the fallacious statement;
- **Label & Context (`LC`)**: we supplied the model with the context, the fallacious statement, and the correct fallacy label;
- **NO Label & Context (`NO`)**: we provided the model with the fallacious statement only, without any additional context or label;
- **Label Only (`LO`)**: we provided the fallacious statement along with its correct label.

When explicit labels are not provided, the model is required to perform a classification task to predict the appropriate fallacy category (see *Fallacy Classification Sub-Task* module in Figure 2).

We evaluated the effectiveness of our approach through both automated metrics (e.g., BERTScore, IOU F1) and human evaluations metrics. In order to qualitatively assess the generated non-fallacious arguments, a rigorous human-in-the-loop evaluation has been addressed along three key dimensions: *Relevance* (i.e., alignment with the original topic), *Suitability* (i.e., appropriateness of the repair), and *Cogency* (i.e., logical coherence and persuasiveness). Seventeen annotators voluntarily evaluated

Models	Techniques											
	Zero-Shot				*Few-Shot*				*Fine-Tuning*			
	CO	LC	NO	LO	CO	LC	NO	LO	CO	LC	NO	LO
BART	-	-	-	-	-	-	-	-	0.98	0.98	-	-
Claude 3	0.68	0.69	0.52	0.54	0.71	**0.78**	0.64	0.65	-	-	-	-
Gemma 1.1 2B	0.62	0.53	0.64	0.49	0.38	0.49	0.39	0.49	0.49	0.49	0.55	0.48
Gemma 1.1 7B	0.55	0.54	0.52	0.50	0.51	0.61	0.49	0.72	0.51	0.51	0.51	0.49
GPT 3.5 turbo	0.68	0.70	0.62	0.59	0.65	-	0.69	0.62	0.68	0.69	0.66	0.63
GPT 4	0.69	**0.71**	0.61	0.60	0.70	-	0.66	-	-	-	-	-
LLaMa 3 8B	0.66	0.67	0.56	0.57	0.70	0.55	0.60	0.52	0.93	0.88	0.96	**0.97**
Mistral 7B	0.58	0.58	0.53	0.53	0.58	0.61	0.58	0.54	-	-	-	-
Mixtral 8x7B	0.60	0.62	0.57	0.53	0.60	0.43	0.59	0.43	0.76	0.79	0.62	0.58

Table 6: Results of BERTScore in all experimental configurations.

15 repaired arguments generated by the top models in each setting and configuration. In terms of ratings, LLM-generated annotations were deemed relevant (4.03 $\pm$ 0.68) and suitable (4.17 $\pm$ 0.68) but were rated lower in Cogency (3.76 $\pm$ 0.69) on a 5-point Likert scale.

Table 6 presents the evaluation results using BERTScore for all tested models on the task of generating non-fallacious arguments using various prompt techniques. Our results demonstrate that LLMs can adequately repair fallacious arguments when guided by targeted prompts or fine-tuned on domain-specific dataset such as the FallacyFix dataset. Emotional appeals (e.g., *Appeals to Fear*) were found to be easier to repair due to their distinct linguistic markers (e.g., exaggerated or dramatic statements Goffredo et al., 2023b), while more complex fallacies (e.g., *Ad Hominem*) required deeper contextual understanding.

While examining the *human evaluation metrics*, we observed a high percentage of agreement between annotators, suggesting that the models often produce content that is fitting and relevant. The analysis also revealed an high percentage of subjectivity in the evaluation, with annotators reaching similar judgement through different reasoning.

The identification of an optimal model for accurate fallacy repair remains a challenging task and depends on the chosen strategy and prompt technique. DISPUTool 3.0 incorporates a *fine-tuned* LLaMA 3 8B (Dubey et al., 2024) in the *Label Only* (LO) setting. Our choice was driven by the results that this model obtained on our benchmark and its significantly lower financial cost compared to other non open-source models.

4 Conclusion

DISPUTool 3.0 is designed for researchers in digital humanities and political communication, and it offers an integrated and modular framework to automatically analyse and assess political debates in English. With respect to the previous version of the tool where argument mining models were employed to identify and classify argument components, some new modules have been included in DISPUTool 3.0. First, the identification of argumentative relations has been improved through the integration of a fine-tuned DeBERTa-V3 model (He et al., 2023), achieving a Macro F1 score of 0.69. This improvement enables a more precise mapping of argumentative structure across complex political debates. Second, DISPUTool 3.0 proposes an automatic fallacy detection and classification module. This functionality leverages the MultiFusion BERT architecture (Goffredo et al., 2023b) reaching a Macro F1 score of 0.74. This new module supports the systematic identification of manipulative or logically flawed arguments within political discourse. Third, DISPUTool 3.0 introduces a *repairing fallacious arguments* module, which automatically generates non-fallacious arguments of the detected fallacious arguments. This generative module is implemented using the LLaMA 3 8B model (Dubey et al., 2024), and represents a step towards counter-narrative generation.

Future research will focus on integrating domain-specific knowledge to address complex fallacy categories, further analyzing language models' behavior in countering fallacies, and exploring real-time fallacy repair methodologies. These efforts aim to enhance our ability to address fallacies dynamically in various argumentation contexts, potentially improving the quality of public discourse and decision-making.

Limitations

Despite the significant advancements presented in DISPUTool 3.0, some limitations have to be discussed: *i)* the tool is trained to analyze political debates in English, which may reduce its performance in non-English speaking contexts; *ii)* while the *ElecDeb60to20* dataset covers US presidential debates, it does not include other forms of debates such as congressional debates, town halls, or international political discussions; *iii)* the process of repairing fallacious arguments involves some degree of subjectivity, meaning that there can be multiple valid ways to formulate a non fallacious version of a fallacious argument. Additionally, this work leverages advanced generative models such as LLaMA 3 8B. Generative models exhibit non-deterministic behavior, producing varied outputs for identical inputs across different instances. This variability may lead to inconsistent or irrelevant outputs. In this work, LLaMA 3 8B was trained over a specific set of fallacies and therefore, it may not work if the fallacious argument we want to repair belongs to a category of fallacies outside this set.

Acknowledgements

This work has been supported by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8172–8187. Association for Computational Linguistics.

Tariq Alhindi, Smaranda Muresan, and Preslav Nakov. 2024. Large language models are few-shot training example generators: A case study in fallacy recognition. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12323–12334. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. 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*, pages 5427–5433. ijcai.org.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2309–2330. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Pierpaolo Goffredo, Elena Cabrio, Serena Villata, Shohreh Haddadan, and Jhonatan Torres Sanchez. 2023a. Disputool 2.0: A modular architecture for multi-layer argumentative analysis of political debates. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 16431–16433. AAAI Press.

Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023b. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11101–11112. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint*

Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, pages 4143–4149. ijcai.org.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian M. Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4810–4845. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

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

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Steve Oswald and Thierry Herman. 2020. Give the standard treatment of fallacies a chance! cognitive and rhetorical insights into fallacy processing. *From Argument Schemes to Argumentative Relations in the Wild: A Variety of Contributions to Argumentation Theory*, pages 41–62.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 644–657. Association for Computational Linguistics.

Frans H. van Eemeren, editor. 2015. *Reasonableness and Effectiveness in Argumentative Discourse, Fifty Contributions to the Development of Pragma-Dialectics*, volume 27 of *Argumentation Library*. Springer.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1519–1538. ELRA and ICCL.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.