Argumentative Fallacy Detection in Political Debates

Eva Cantín Larumbe and Adriana Chust Vendrell
Universitat Politècnica de València, Spain
{ecanlar, achuven}@etsinf.upv.es

Abstract

Building on recent advances in Natural Language Processing (NLP), this work addresses the task of fallacy detection in political debates using a multimodal approach combining text and audio, as well as text-only and audio-only approaches. Although the multimodal setup is novel, results show that text-based models consistently outperform both audio-only and multimodal models, confirming that textual information remains the most effective for this task. Transformer-based and few-shot architectures were used to detect fallacies. While fine-tuned language models demonstrate strong performance, challenges such as data imbalance, audio processing, and limited dataset size persist.

1 Introduction

Recent advances in Natural Language Processing (NLP) have enabled substantial progress in understanding and generating human language. Within this context, fallacy detection has gained attention as a means to promote critical thinking and mitigate the spread of misinformation. Fallacious arguments—persuasive yet logically flawed—can contribute to the circulation of deceptive content, especially in politically charged discourse.

Automatic fallacy detection in political discourse could support content moderation, enhance public debate quality, and foster critical thinking by identifying manipulative rhetorical strategies at scale.

Fallacies have been studied extensively in argumentation theory. Informally, they are arguments that appear sound but contain subtle logical flaws (Breslin, 2023). Walton (2006) defines a fallacy as "an argument that seems valid on the surface but is flawed when examined more closely."

In this paper, we address fallacy detection (AFD) in the domain of political debates. This work is part of the MM-ArgFallacy2025 Shared Task on Multimodal Argumentative Fallacy Detection and Classification on Political Debates, co-located with the 12th Workshop on Argument Mining in Vienna, Austria.

We investigate the performance of Transformer-based and few-shot models, across three input configurations: text-only, audio-only, and a novel multimodal setting that combines both modalities. Our contributions are threefold: (1) we apply argumentation theory to fallacy detection in political discourse; (2) we evaluate and compare the performance of text-only, audio-only, and multimodal Transformer-based architectures; and (3) we analyze key challenges in fallacy detection, including data imbalance, the limited contribution of audio features, and practical constraints related to computational resources and training time.

2 Related Work

Recent advances in fallacy detection span from supervised learning with curated datasets to zero-shot prompting with Large Language Models (LLMs), as well as emerging multimodal approaches.

Chaves et al. (2025) introduced FALCON, a multi-label, graph-based dataset focused on fallacies in COVID-19 and politically charged discourse on Twitter. Annotated by experts across six fallacy types, the dataset supports multiple labels per instance. Among the models evaluated, a dual-transformer architecture augmented with sentiment scores and contextual cues achieved the best performance, with a macro F1 score of 48.8%.

Similarly, Atarama et al. (2024) developed F-Detector, a BERT-based classifier for detecting ten fallacy types in digital texts. Leveraging advanced NLP preprocessing and dataset expansion via generative techniques, their system attained a substantial improvement over previous models—achieving an F1 score of 74%, significantly outperforming GPT-based baselines.

Moving toward generalization, Pan et al. (2024)

Proceedings of the 12th Argument Mining Workshop, pages 369–373

explored the use of LLMs as zero-shot fallacy classifiers, mitigating the reliance on annotated datasets. They proposed single-round and multi-round prompting strategies to enhance fallacy reasoning. Evaluating models like GPT-4 across seven benchmark datasets—including political debates and COVID-19 discourse—they found that LLMs outperformed fine-tuned models (e.g., T5) in out-of-distribution (OOD) settings, especially when multi-round prompting was used to aid smaller models.

In the specific domain of political discourse, Cruz et al. (2025) introduced the FallacyES-Political dataset, comprising 1,965 annotated fallacies from 30 years of Spanish electoral debates. Their study compared zero-shot GPT-4o and a fine-tuned RoBERTa-base-BNE model, with the latter achieving superior performance (F1 score of 0.641 vs. 0.570), underscoring the value of domain-specific fine-tuning on curated data.

Finally, in a novel direction, Mancini et al. (2024) proposed a multimodal framework for fallacy classification, introducing MM-USED-fallacy, the first dataset combining text and audio from U.S. presidential debates. Their experiments showed that integrating audio features—especially for fallacies like Appeal to Emotion and Appeal to Authority—can yield significant gains, with multimodal models outperforming text-only baselines by up to 8 percentage points in F1 score.

In contrast to previous approaches, our work is among the first to address the detection of fallacies in political debates using a multimodal framework that integrates both textual and audio features.

3 Data

We utilized the MM-USED-fallacy dataset (Mancini et al., 2024), which includes 17,118 instances, with 15,550 labeled as non-fallacious and 1,568 containing a fallacy. The test set comprises 2,175 samples.

The task was approached using three distinct data configurations: text-only, audio-only, and multimodal (text + audio). This setup allowed for an evaluation of the performance of each modality individually, as well as an exploration of the potential synergies from combining text and audio information.

In our methodology, we did not incorporate contextual information in the experiments. This was primarily due to constraints in available computational resources and limited time, which prevented us from implementing and testing models that consider extended discourse context.

For the text modality, no additional preprocessing was applied beyond tokenization using the pre-trained model's tokenize. For the audio modality, two approaches were used. In the mel-spectrogram + CNN setup, audio was loaded at 22,050 Hz, trimmed or zero-padded to 3 seconds, converted to 128-band mel spectrograms (fmax=8,000 Hz), normalized to [0, 255], resized to 128×128 pixels, and stacked into 3 channels. In the Wav2Vec 2.0 setup, raw audio was loaded at 16,000 Hz and processed using the facebook/wav2vec2-base processor, which handled feature extraction and padding. For the combined text-audio setting, we adopted a frozen RoBERTa (base) encoder for text and Wav2Vec 2.0 for audio. The text data were tokenized using the RobertaTokenizer, and the audio inputs were processed with the Wav2Vec2Processor.

Finally, data was split consistently. The training and validation sets were obtained using an 80/20 stratified split to maintain label distribution.

4 Experiments

4.1 Method

To build upon previous experimental findings in the literature, we implemented a series of deep learning models. Specifically, we explored five different models for the text-only modality: BERT (uncased) (Devlin et al., 2018), RoBERTa (base and large) (Liu et al., 2021), SBERT (Reimers and Gurevych, 2019), ALBERT (Base v2) (Lan et al., 2019), and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025).

For BERT and ALBERT, we unfroze the last two hidden layers of the encoder in addition to the classification layer. In the case of RoBERTa and SBERT, we unfroze the last four hidden layers. Regarding DeepSeek-R1-Distill-Llama-8B, we loaded the model with 4-bit quantization using the NF4 scheme and bfloat16 computation, following QLoRA best practices. This quantization approach substantially reduced memory usage and computational overhead, enabling efficient fine-tuning of large-scale models on consumer-grade hardware. Despite the reduced precision, the NF4 scheme preserved high performance by employing a non-uniform quantization grid optimized for downstream tasks.

For the fallacy detection task, the prompt designed for DeepSeek-R1-Distill-Llama-8B was inspired by Ruiz-Dolz and Lawrence (2023):

```
Your task is to detect the type
of fallacy in the Text. The label
should be 1 (it is a fallacy) or
0 (it is not a fallacy)
Text Snippet: [SAMPLE]
```

For the audio-only approach, we tested two models: MFCC + CNN and Wav2Vec 2.0 (Schneider et al., 2019). The MFCC + CNN model uses handcrafted audio features, while Wav2Vec 2.0 processes raw audio waveforms with a pretrained deep learning model. This comparison helps evaluate traditional feature-based methods versus end-to-end representation learning for fallacy detection.

For the text-audio approach, we used a combined model of RoBERTa (base) (Liu et al., 2021) and Wav2Vec2-Base-960h (Schneider et al., 2019). We selected this combination because RoBERTa and Wav2Vec2-Base-960h were the best-performing models in the text-only and audio-only settings, respectively, providing a strong foundation for the multimodal setup. Features from both encoders were concatenated and passed through a classification head. To reduce computational cost, only the classification layers were unfrozen.

4.2 Experimental Setup

Hyperparameter selection was performed based on validation performance. The best model per configuration was retrained during 3-5 epochs on the full training set and evaluated on the test set. Finally, this retrained model was used to generate predictions on the test set.

Text-Only Models. We experimented with five Transformer-based models: BERT (uncased), RoBERTa (base and large), SBERT, ALBERT (Base v2), and DeepSeek-R1-Distill-Llama-8B. For BERT, class imbalance was addressed using weighted loss: we used weights of 0.2 for the non-fallacious class and 0.8 for the fallacious class. These weights were selected empirically based on preliminary validation performance, aiming to improve the F1-score for the minority class. This approach allowed us to balance sensitivity to both classes during optimization.

All models, except for DeepSeek, were trained using a learning rate of 2e-5, cosine learning rate scheduling, a weight decay of 0.1, and 10 epochs. For DeepSeek, a linear classification head was added on top of mean-pooled hidden states, and the prompt used is described in the Method Section 4.1.

Audio-Only Models. For audio-only models, raw audio samples were converted into 128×128 mel-spectrogram images, which were then used to train a CNN composed of three convolutional blocks followed by max-pooling, dropout, and dense layers. The CNN was optimized for binary classification using the Adam optimizer. In parallel, a Wav2Vec2.0-based model was fine-tuned to perform classification directly from raw audio waveforms. The audio files were first loaded and preprocessed using a pre-trained Wav2Vec2 processor, which performed feature extraction and normalization. The extracted features were then passed to a Wav2Vec2 model with a classification head adapted for binary classification.

Multimodal Model. The model was trained for 10 epochs using a class-weighted Cross-Entropy loss with weights of 50 for class 1 and 1 for class 0. These weights were selected empirically based on preliminary validation experiments to better handle class imbalance. Optimization was performed using AdamW and a ReduceLROnPlateau scheduler with an initial learning rate of 2e-5.

5 Results

5.1 Validation Results

All performance metrics were computed on the validation set. We report Accuracy (Acc.) and Binary F1-score. Results for the task are presented in Table 1.

As shown in Table 1, the best performance was obtained by RoBERTa in the text-only setting, achieving the highest Binary F1-score. Other Transformer-based models performed similarly, while the DeepSeek-R1 zero-shot model lagged considerably. This can be attributed to its zero-shot nature: unlike models like RoBERTa or BERT, which were fine-tuned on the task-specific data, DeepSeek-R1 was evaluated without any additional training. Since fallacy detection requires nuanced, context-aware understanding of argumentative language, zero-shot models often fail to capture task-specific patterns, resulting in lower performance. Fine-tuning DeepSeek-R1 on task-specific fallacy detection data could significantly improve its performance.

Model	Acc.	Binary F1
Text-only		
BERT	0.9042	0.3037
RoBERTa	0.9033	**0.3393**
ALBERT	0.9004	0.2816
SBERT	0.9077	0.2956
DeepSeek-R1	0.8814	0.1567
Audio-only		
MFCC+CNN	0.3902	0.1618
Wav2Vec2	0.0938	**0.1683**
Text-audio		
RoBERTa+Wav2Vec2	0.4866	**0.1831**

Table 1: Accuracy and Binary F1-score for the fallacy detection task (validation set).

Audio-only models yielded substantially lower performance across both metrics, with Wav2Vec 2.0 slightly outperforming the CNN-based approach. This limited effectiveness of audio features may be due to several factors: first, fallacy detection primarily relies on semantic and contextual understanding, which is inherently stronger in textual data than in acoustic signals. Second, the acoustic cues relevant to detecting fallacies—such as tone, emphasis, or hesitation—might be too subtle or inconsistent to be reliably captured by current audio representations.

The multimodal configuration (RoBERTa + Wav2Vec 2.0) showed marginal improvement over audio-only models but remained well below the performance of text-only models. These results suggest that semantic cues in text are more informative for fallacy detection, while the additional acoustic features did not contribute significantly under the current setup.

5.2 Official Test Set Results and Shared Task Ranking

Table 2 presents the performance of our best model on the official test set provided by the MM-ArgFallacy2025 shared task. These results reflect the final evaluation submitted to the organizers and were used to determine our ranking in the competition.

Modality	Model	Binary F1	Ranking
Text-only	RoBERTa	0.2195	4th
Audio-only	Wav2Vec2	0.1690	2nd
Text-Audio	RoBERTa + Wav2Vec2	0.1931	4th

Table 2: Binary F1-score and ranking on the official test set for the fallacy detection task, grouped by modality.

6 Conclusion

This work presented a comprehensive evaluation of deep learning models for the detection of logical fallacies in political debates, leveraging both text and audio modalities. Our best-performing model—a fine-tuned RoBERTa variant—achieved an accuracy of 90.33% and a binary F1 score of 0.3393.

Our findings support and extend previous work such as Mancini et al. (2024), while offering new insights. Unlike prior approaches that emphasize multimodal fusion, our experiments indicate that text-only models consistently outperform audio-only and multimodal models for both tasks. In particular, RoBERTa achieved the highest score, underscoring the strength of contextualized language representations in reasoning-based classification tasks.

The proposed model has the potential to be deployed in various applications, such as automatic detection of fallacious reasoning in online forums, academic writing, or news articles. This could aid in improving the quality of discourse in these environments by flagging problematic arguments. Additionally, educational tools could benefit from such a model to help students learn to identify and avoid common logical fallacies in their reasoning.

The code, trained models, and detailed experimental results presented in this work are publicly available at our GitHub repository, facilitating reproducibility and further research in fallacy detection.

7 Limitations

While our proposed model achieved promising results for fallacy detection, several limitations must be acknowledged. The dataset was highly imbalanced, with a significantly larger number of non-fallacious examples compared to fallacious ones. This imbalance likely impacted the model's ability to generalize effectively across both classes, causing it to be biased toward the majority non-fallacious category.

Computational constraints also posed significant challenges. The majority of experiments were conducted using limited GPU resources—dual T4

GPUs on Kaggle (restricted to 30 hours per week) and a local RTX 4070 setup. These limitations prevented thorough hyperparameter tuning and restricted the number of training epochs, particularly for computationally intensive models such as Wav2Vec2 and multimodal architectures. Audio-only and text-audio models were disproportionately affected, as their training was slower and more resource-intensive.

Additionally, time constraints further limited the breadth of our experimentation. In some cases, a single training epoch required up to 30 minutes, significantly curtailing our ability to explore alternative architectures and training strategies. As a result, the full potential of multimodal learning in this context remains underexplored.

These limitations are consistent with broader challenges reported in the field of fallacy detection. Many prior studies also rely on small or imbalanced datasets, limiting generalizability across fallacy detection or application domains. Model performance tends to vary significantly depending on the modality used—text, audio, or multimodal—which complicates cross-study comparisons.

Future work should address these issues by expanding and balancing the dataset across binary fallacy detection categories, optimizing training efficiency, and leveraging more robust computational infrastructure. Exploring a wider range of multimodal architectures with better scalability would also be essential for capturing nuanced fallacious patterns beyond textual content alone. Additionally, an ablation study that systematically repeats all experimental settings while incorporating contextual information—such as preceding or surrounding sentences—could help quantify the impact of context on fallacy detection performance and better inform future model designs.

References

Diego Atarama, Diego Pereira, and Cesar Salas. 2024. F-detector: Design of a solution based on machine learning to detect logical fallacias on digital texts. In *2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 216–221. IEEE.

Frank Breslin. 2023. Fallacy Detection: Part 1 — frankbreslin41. https://medium.com/@frankbreslin41/fallacy-detection-part-1-2e5047c335b9.

Mariana Chaves, Elena Cabrio, and Serena Villata. 2025. Falcon: A multi-label graph-based dataset for fallacy classification in the covid-19 infodemic. In *SAC'25-ACM/SIGAPP Symposium on Applied Computing*.

Fermín L Cruz, Fernando Enríquez, F Javier Ortega, and José A Troyano. 2025. Fallacyes-political: A multiclass dataset of fallacies in spanish political debates. *Procesamiento del Lenguaje Natural*, 74:127–138.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *China national conference on Chinese computational linguistics*, pages 471–484. Springer.

Eleonora Mancini, Federico Ruggeri, Paolo Torroni, and 1 others. 2024. Multimodal fallacy classification in political debates. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 170–178. Association for Computational Linguistics.

Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. 2024. Are llms good zero-shot fallacy classifiers? *arXiv preprint arXiv:2410.15050*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Douglas Walton. 2006. *Fundamentals of Critical Argumentation*. Cambridge University Press, New York.