Multimodal Argumentative Fallacy Classification in Political Debates

Warale Avinash Kalyan, Siddharth Pagaria, Chaitra V, Spoorthi H G
Presidency University, Bengaluru, India
warale.avinash@gmail.com

Abstract

Argumentative fallacy classification plays a crucial role in improving discourse quality by identifying flawed reasoning that may mislead or manipulate audiences. While traditional approaches have primarily relied on textual analysis, they often overlook paralinguistic cues such as intonation and prosody that are present in speech. In this study, we explore how multimodal analysis, in which we combine textual and audio features, can enhance fallacy classification in political debates. We develop and evaluate text-only, audio-only, and multimodal models using the MM-USED-fallacy dataset to assess the contribution of each modality. Our findings indicate that the multimodal model, which integrates linguistic and acoustic signals, outperforms unimodal systems, underscoring the potential of multimodal approaches in capturing complex argumentative structures.

1 Introduction

Argumentative fallacies are the reasoning errors that may appear rhetorically persuasive yet lack logical validity. They pose a significant challenge to both critical thinking and automated discourse analysis. In high-stakes communicative contexts such as political debates, these fallacies (e.g., *ad hominem, appeal to emotion, slippery slope, false cause*) are frequently employed to sway audiences while circumventing sound logic. Automatically identifying such flawed reasoning patterns, a task known as argumentative fallacy classification, is increasingly recognized as a crucial objective in computational argumentation with implications for misinformation detection, media literacy, and democratic accountability.

Recent work has demonstrated the potential of large-scale pretrained language models for detecting fallacies in text. Jin et al. (2022) introduced a benchmark taxonomy and showed that transformer-based models such as RoBERTa outperform traditional classifiers. Goffredo et al. (2022) extended this research to political discourse, annotating U.S. presidential debates and highlighting the importance of nuanced semantic understanding for identifying reasoning flaws. These contributions underscore the ability of neural models to capture structural properties of argumentation when grounded in high-quality text data.

Fallacious reasoning often depends not only on what is said but also on how it is delivered. Paralinguistic features such as intonation, stress, and rhythm convey speaker intent and emotional appeal. Early multimodal work like M-Arg (Mestre et al., 2021) combined audio and transcripts to enhance argumentative analysis, while (Mancini et al., 2022) showed how prosodic signals complement lexical cues in detecting fallacies.

Building on this, Mancini et al. (2024a) introduced MAMKit, which includes the MM-USED-fallacy dataset (Mancini et al., 2024b), annotated with aligned audio and text across six fallacy types. In this paper, we evaluate a broad set of models and focus on three best-performing configurations based on validation performance: text-only (RoBERTa), audio-only (BiLSTM with MFCC), and multimodal (RoBERTa with Wav2Vec2), assessed under a unified framework for comparative analysis. [1]

2 Related Work

The classification of argumentative fallacies has evolved from early rule-based and shallow learning methods to modern neural architectures built on large-scale pre-trained language models. Jin et al. (2022) framed fallacy detection as a structured classification task and demonstrated the advantages of transformer-based approaches, such as RoBERTa, in capturing complex reasoning patterns. Goffredo et al. (2022) extended this line of work to political

[1] https://github.com/p4rz1v4l26/Multimodal-Argumentative-Fallacy-Classification-in-Political-Debates

Proceedings of the 12th Argument Mining Workshop, pages 374–380

discourse, introducing a richly annotated corpus of U.S. presidential debates and showing that encoding argument structure improves textual fallacy classification.

Beyond text, multimodal approaches have gained traction as researchers increasingly recognize the role of delivery in persuasive discourse. Mestre et al. (2021) introduced M-Arg, a dataset that combines transcripts and aligned audio from political debates, showing that models incorporating both modalities outperform unimodal baselines. Mancini et al. (2024b) released the MM-USED-fallacy corpus, which includes six fallacy categories annotated over real-world political debate clips. This was followed by the release of MAMKit (Mancini et al., 2024a), a toolkit that provides standardized preprocessing and modeling routines for this dataset. Their work highlighted how prosodic cues can complement lexical signals in fallacy detection.

While prior studies highlight the potential of multimodal approaches, they often lack systematic comparisons across modalities. In our work, we evaluate several transformer-based text models and audio models, ultimately selecting RoBERTa for text and BiLSTM for audio based on validation performance. For the multimodal setup, we combined RoBERTa with Wav2Vec2.0. These three configurations were chosen for their strong performance under consistent settings on the MM-USED-fallacy dataset, forming the basis of our controlled comparison across the modalities.

3 Data

We performed the experiments on the **MM-USED-fallacy** dataset (Mancini et al., 2024b), a multimodal resource released as part of the MAMKit toolkit for argument mining. This dataset is specifically designed for the **Argumentative Fallacy Classification (AFC)** task and contains aligned textual and audio segments drawn from political debates. Each snippet is annotated with one of six fallacy types: *ad hominem*, *appeal to authority*, *appeal to emotion*, *slippery slope*, *slogans*, and *false cause*.

Inspired by the setup in Mancini et al. (2024b), our work leverages both linguistic and paralinguistic information from the MM-USED-fallacy dataset. Table 1 presents the count of instances for each fallacy type. This distribution provides insight into the prevalence of each class within the dataset and informs model training, particularly in terms of addressing class imbalance. Notably, some categories such as *appeal to emotion* and *ad hominem* occur more frequently, whereas others like false cause and slogans are relatively underrepresented, potentially impacting classification performance.

Fallacy	**MM-USED-fallacy**
Appeal to Emotion	800
Appeal to Authority	191
Ad Hominem	149
False Cause	56
Slippery Slope	46
Slogans	36
Total Count	**1,278**

Table 1: Distribution of fallacy types in the MM-USED-fallacy dataset.

We employed a stratified data splitting strategy using the `mm-argfallacy-2025` custom dataset splitter, introduced (Mancini et al., 2024a) as part of the MAMKit toolkit. This splitter partitions the data into non-overlapping train, validation, and test sets while maintaining label distribution. The final evaluations were conducted on a held-out secret test set to ensure unbiased assessment of model performance. For further details refer Appendix B.

3.1 Preprocessing and Cleaning

The preprocessing pipeline was tailored to meet the requirements of unimodal and multimodal classification models:

Text Modality.

BERT Text was tokenized using the `BertTokenizer`. Inputs were lowercased (for `bert-base-uncased`), tokenized using WordPiece encoding, and padded or truncated to a fixed sequence length.

RoBERTa We used the `RobertaTokenizer` from Hugging Face. To incorporate broader context, each sentence was concatenated with its preceding and following sentences. Standard text normalization procedures were applied to eliminate inconsistencies, special characters, and formatting noise.

DeBERTa The `DebertaTokenizer` was used for tokenization. Similar to RoBERTa, preprocessing included sentence normalization and cleaning. The

pipeline was adapted to accommodate DeBERTa's disentangled attention mechanism.

Audio-Modality.

BiLSTM + MFCC Audio recordings were converted to mono-channel at 16 kHz and standardized to a duration of 5 seconds via padding or truncation. We extracted 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) using Librosa, followed by mean-variance normalization to stabilize training.

Wav2Vec2 Raw audio waveforms sampled at 16 kHz were fed directly into the `wav2vec2-base-960h` model without handcrafted feature extraction. Padding or truncation was applied to conform to model input constraints.

Text-Audio Modality.

RoBERTa + Wav2Vec2 Text and audio inputs were preprocessed independently, following the procedures described in the respective unimodal sections. Text was tokenized using the `RobertaTokenizer`, with adjacent sentences concatenated to provide contextual information. Audio inputs were raw waveforms sampled at 16 kHz and padded or truncated to a fixed length of 5 seconds before being passed to the `wav2vec2-base-960h` model. This ensured consistency in input dimensions across both modalities.

4 Experimental Setup

This section outlines the overall architecture and training configuration of models developed for argumentative fallacy classification using text, audio, and multimodal inputs. The models are evaluated using the MM-USED-fallacy dataset, which comprises annotated conversational data collected from political discourse. As illustrated in Figure 1, the multimodal framework integrates a text module and an audio module, whose respective feature representations are concatenated and passed through a classifier to predict the fallacy label.

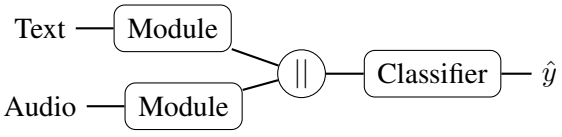

Figure 1: The schema for multimodal Argumentative Fallacy Classification model.

4.1 Model

We evaluated three distinct model configurations for fallacy classification: a Text-Only Model, an Audio-Only Model, and a Text-Audio Model. Each model is trained independently and assessed on the validation dataset to enable comparative analysis. Based on the results achieved and displayed in Table 2, we elected to proceed with the model that demonstrated the highest F1-score across labels on the validation set.

Model	AH	AE	AA	FC	SS	S	Average ($\bar{x} \pm \sigma$)
Text-only							
RoBERTa	**.10**	**.81**	**.22**	**.19**	**.11**	**.03**	**.24 ± .26**
BERT	.09	.74	.2	.16	.12	.02	.22 ± .13
DeBERTa	.06	.13	.12	.08	.07	.01	.078 ± .08
Audio-only							
BiLSTM w/ MFCC	**.00**	**.76**	**.05**	**.11**	**.06**	**.00**	**.16 ± .38**
Wav2Vec2	.00	.56	.04	.11	.05	.00	.12 ± .06
Multi-Modal							
RoBERTa + Wav2Vec2	**.09**	**.79**	**.19**	**.09**	**.07**	**.06**	**.22 ± .27**

Table 2: Macro F1-scores across fallacy types for each model configuration. AH: Ad Hominem, AE: Appeal to Emotion, AA: Appeal to Authority, FC: False Cause, SS: Slippery Slope, S: Slogans.

4.1.1 Text-Only Model

The text-only models are trained to classify fallacies using only the linguistic content of annotated snippets. We experiment with three transformer-based architectures: DeBERTa, BERT, and RoBERTa, each trained on the fallacy-labeled text segments. These models enable a comparative analysis of how different pretrained language encoders capture argumentative patterns in political discourse. The results presented in Table 2 are based on validation data and reflect the performance of the models under a standardized training setup. As seen in Table 2, RoBERTa and BERT outperform the other models, achieving the highest validation F1-score. Based on this observation, we selected RoBERTa as the final text encoder for our text-only and multimodal configurations due to its consistent performance.

4.1.2 Audio-Only Model

We evaluated two audio-only pipelines: one using MFCC features with a BiLSTM classifier, and another using raw audio with a pretrained Wav2Vec2 encoder. In the MFCC-BiLSTM setup, audio clips were converted to 16 kHz mono and standardized to 5 seconds by padding or truncation. We extracted 13-dimensional MFCC features using Librosa, which capture tone and rhythm patterns, and

fed them into a BiLSTM for temporal modeling, followed by a dense classification layer (Aldeneh and Provost, 2017). The Wav2Vec2 pipeline, by contrast, operated directly on raw audio to extract high-level embeddings. As shown in Table 2, the MFCC-BiLSTM model outperformed Wav2Vec2 on the validation set and was selected for further experimentation.

4.1.3 Text-Audio Model

The multimodal architecture integrates both textual and audio modalities to enhance fallacy detection performance. For the textual modality, we employ a pre-trained RoBERTa model as the unimodal text encoder, extracting contextual embeddings from input sequences. For the audio modality, we utilize Wav2Vec2 to encode raw audio signals into high-level feature representations. The outputs from both unimodal encoders are then concatenated and fed into a logistic regression meta-classifier, which performs the final classification. This late fusion strategy allows the model to leverage complementary information from both text and audio streams, facilitating more robust fallacy identification. The validation F1-score of the text-audio model is shown in Table 2.

4.2 Model Training

Model training was conducted under constrained computational resources, without access to a dedicated GPU. This limitation imposed significant restrictions on batch size, model complexity, and training time, thereby influencing design choices throughout our experiments. Due to these software and hardware constraints, lightweight architectures and efficient preprocessing pipelines were prioritized. Kindly refer to Appendix A for more details on training configuration and hyperparameter settings.

4.3 Role of the Meta-Classifier

For the multimodal pipeline, we adopted a late fusion strategy, where a logistic regression *meta-classifier* combines the feature representations from the unimodal text and audio encoders. While this approach allows aggregation of complementary representations, its benefits were limited under current conditions, likely due to weak individual model confidence on rare classes and high modality noise. Future work could explore deeper fusion strategies to improve effectiveness.

5 Results

We evaluated three distinct configurations for the task of argumentative fallacy classification in political debates: a text-only model, an audio-only model, and a multimodal text-audio model. Model performance was assessed on the test set. Table 3 shows the macro F1 scores values of our proposed models, alongside results from other participating teams in the shared task, enabling a direct comparison of system performances, alongside their respective baselines.

Team Name	F1-Score
Text-only	
Team NUST	0.4856
Baseline BiLSTM	0.4721
Alessiopittiglio	0.4444
Baseline RoBERTa	0.3925
Team EvaAdriana	0.3746
Team CASS	**0.1432**
Audio-only	
Alessiopittiglio	0.3559
Team EvaAdriana	0.1858
Team NUST	0.1588
Baseline BiLSTM + MFCC	0.1582
Team CASS	**0.0864**
Baseline WavLM	0.0643
Text-Audio	
Team NUST	0.4611
Alessiopittiglio	0.4403
Baseline RoBERTa + WavLM	0.3816
Team EvaAdriana	0.3746
Baseline BiLSTM + MFCC	0.2191
Team CASS	**0.1432**

Table 3: Performance (F1-score) of our models (Team CASS) on the shared task test set, compared with other participating systems and official baselines

Overall, the classification results reveal relatively low performance across all models, with macro-F1 scores ranging from 0.08 to 0.14 (Table 3). While the audio-only model produced slightly different results compared to the text-only and multimodal configurations, it exhibited a significantly lower F1-score, indicating imbalanced precision and recall across classes. This may hinder consistent fallacy classification performance, especially in the presence of class imbalance.

5.1 Analysis of Results

These outcomes suggest that **textual cues remain the most reliable modality** in fallacy classification, aligning with findings from Jin et al. (2022) and Mancini et al. (2024b). Despite employing pretrained architectures for both text and audio modalities (Mancini et al., 2024b), our models exhibited relatively low macro-F1 scores across all configurations. This underperformance, detailed in Tables 2 and 3, is not merely an artifact of architecture selection but reflects deeper challenges inherent in the dataset and experimental constraints. Factors that may contribute to this are as follows:

Overfitting and Generalization Failure. We observe a significant discrepancy between validation and test performance, largely due to overfitting. As shown in Table 2, models achieve high F1-scores for the dominant class *Appeal to Emotion* (over 70%), and fail to generalize fallacy types, which constitutes the majority of both validation and test data. Consequently, when the test distribution slightly shifts or includes more ambiguous examples, performance drops sharply. This overfitting is likely exacerbated by severe class imbalance, which causes the model to memorize rather than learn fallacy-specific patterns.

Class imbalance and Limited Training. As shown in Table 1, the MM-USED-fallacy dataset is heavily skewed towards "Appeal to Emotion," which comprises over 60% of the samples. This imbalance likely biases model predictions toward dominant classes and penalizes underrepresented ones like "Slogans" or "Slippery Slope." The models were trained under constrained computational settings, with only 3–5 training epochs per configuration. In contrast, prior baselines, such as those reported in Mancini et al. (2024b) were trained for up to 500 epochs. Kindly refer to Appendix A for more details on training configuration.

Multimodal misalignment. Although the dataset contains aligned audio and text, the quality of alignment can vary. Minor temporal mismatches or noisy segments may hinder the effectiveness of Wav2Vec2 embeddings, especially when combined with textual representations.

Limited dataset size. With only 1,278 samples and significant class disparity, models especially with deep architectures like RoBERTa and Wav2Vec2, may be prone to overfitting or undergeneralization.

5.2 Label-Wise Performance

Detailed class-wise performance (Table 2) further confirms that models struggle to predict minority classes. For example, "Slogans" and "Slippery Slope" received near-zero F1 scores across all models, while "Appeal to Emotion" showed high F1 scores. Table 3 reports the macro f1-score for each fallacy category, averaged across all models. These scores reflect model performance on the validation set and illustrate the impact of class imbalance on model behavior.

5.3 Data and Alignment.

During preprocessing, we identified instances of misaligned or corrupted audio-text pairs, similar to the alignment issues noted by Mancini et al. (2024b). One notable case involved the audio file `653.wav` under the dialogue folder 46_2020, which was found to be corrupted and unreadable. According to the dataset, this sample was labeled as *Appeal to Emotion*, and the corresponding dialogue was the phrase *"Excuse me"*. Due to the corrupted audio and the impossibility of establishing a valid alignment, we excluded this sample from our corpus. This exclusion was part of a broader quality control effort aimed at ensuring the reliability of audio-text pairs used in our unimodal and multimodal models. Model performance is influenced by the quality of text-audio alignment. Imperfect or noisy alignments can lead to incomplete multimodal inputs, negatively affecting classification accuracy.

6 Conclusion

This study underscores the enduring primacy of textual semantics in argumentative fallacy classification, while also illuminating the potential and current limitations of multimodal integration. Despite modest gains, the multimodal model's performance reveals unresolved challenges in aligning linguistic and acoustic signals, particularly under class imbalance and data sparsity (Mancini et al., 2024b; Mestre et al., 2021). These findings call for deeper representational synergy across modalities and more robust, corpora rich in argumentative discourse to advance the frontier of computational argumentation in real-world settings.

References

Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition.

In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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-22)*.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

Eleonora Mancini, Federico Ruggeri, Stefano Colamonaco, Andrea Zecca, Samuele Marro, and Paolo Torroni. 2024a. MAMKit: A comprehensive multimodal argument mining toolkit. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)* pages, Bangkok, Thailand. Association for Computational Linguistics.

Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torroni. 2022. Multimodal argument mining: A case study in political debates. In *Proceedings of the 9th Workshop on Argument Mining*.

Eleonora Mancini, Federico Ruggeri, and Paolo Torroni. 2024b. Multimodal fallacy classification in political debates. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers*.

Rafael Mestre, Razvan Milicin, Stuart E Middleton, Matt Ryan, Jiatong Zhu, and Timothy J Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*.

A Training Details

This appendix outlines the implementation framework and experimental configurations used to conduct our study on multimodal argumentative fallacy classification. All experiments were conducted on a system equipped with an Intel Core i5 processor and 8 GB of RAM. The system used an integrated Intel Iris Xe graphics card, which handled all computational tasks during model training and inference. Each model required approximately 6 hours to complete training.

Despite the absence of a dedicated GPU, the experiments were optimized to run efficiently within these hardware constraints. The following tables present the detailed hyperparameter configurations used across our experiments:

Modality	Model	Ep.	BS	LR
Text	RoBERTa	3	8	2e–5
Audio	BiLSTM + MFCC	5	8	1e–3
Text+Audio	RoBERTa + Wav2Vec2	5	16	2e–5

Table 4: Hyperparameters used for each model. Ep: Epochs, BS: Batch Size, LR: Learning Rate.

B Data Loading

To facilitate standardized experimentation, we adopted the data loading and splitting utilities introduced by (Mancini et al., 2024a) for the MM-USED-fallacy dataset, targeting the task of Argumentative Fallacy Classification (AFC). The loader initializes the dataset with the task parameter set to 'AFC'. For consistency in evaluation, we utilize the custom dataset split defined as `mm-argfallacy-2025`, accessed through the get_splits() method. This splitter provides a 70:15:15 ratio for training, validation, and test sets, ensuring dialogue-level separation to prevent context leakage. The use of this academically validated split facilitates meaningful comparisons with prior work. By leveraging this modular and well-supported pipeline, we ensure that our experiments conform to the dataset's structure and are directly comparable with established baselines in the field.