Prompt-Guided Augmentation and Multi-modal Fusion for Argumentative Fallacy Classification in Political Debates

Abdullah Tahir[*], **Imaan Ibrar**[*], **Huma Ameer**[*], **Mehwish Fatima**[*] **and Seemab Latif**[†]
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology, Islamabad, Pakistan
{abtahir.bese21seecs, iibrar.bese21seecs, hameer.msds20seecs,
mehwish.fatima, seemab.latif}@seecs.edu.pk

Abstract

Classifying argumentative fallacies in political discourse is challenging due to their subtle, persuasive nature across text and speech. In our MM-ArgFallacy Shared Task submission, Team NUST investigates uni-modal (text/audio) and multi-modal (text+audio) setups using pretrained models—RoBERTa for text and Whisper for audio. To tackle severe class imbalance, we introduce Prompt-Guided Few-Shot Augmentation (PG-FSA) to generate synthetic samples for underrepresented fallacies. We further propose a late fusion architecture combining linguistic and paralinguistic cues, enhanced with balancing techniques like SMOTE and Focal Loss. Our approach achieves top performance across modalities, ranking 1st in text-only and multi-modal tracks, and 3rd in audio-only, on the official leaderboard. These results underscore the effectiveness of targeted augmentation and modular fusion in multi-modal fallacy classification.

1 Introduction

Argumentative fallacies—reasoning patterns that appear logically sound but are actually flawed—are frequently employed in political discourse to mislead audiences and manipulate opinions (Goffredo et al., 2022). Their subtle persuasive nature can distort public perception and potentially lead to misguided policy decisions. As political debates continue to be a major platform for shaping public opinion, the automatic detection and classification of such fallacies is crucial for fostering transparency and informed democratic dialogue.

While prior work has focused predominantly on textual data using transformer-based models like BERT and RoBERTa (Goffredo et al., 2022, 2023), fallacies are not purely linguistic. Paralinguistic cues such as intonation, pitch, rhythm, hesitation are critical in signaling fallacy types, especially in speech. Emotional appeals and ad hominem attacks often rely heavily on such acoustic features (Mancini et al., 2024b). This motivates a multi-modal perspective for fallacy detection.

Fallacy Type	Description
Ad Hominem	Personal attacks instead of addressing the argument.
Appeal to Authority	Unjustified reliance on authority as evidence.
Appeal to Emotion	Persuasion by emotional manipulation rather than logic.
False Cause	Incorrect causal attributions without sufficient evidence.
Slogan	Use of catchphrases lacking argumentative substance.
Slippery Slope	Assuming one action leads to extreme outcomes without basis.

Table 1: Macro-level argumentative fallacy types and their descriptions (Goffredo et al., 2022).

To address these challenges, the *12th Workshop on Argument Mining* introduces the MM-ArgFallacy Shared Task[1], targeting fallacy detection and classification in political debates under three input settings: text-only, audio-only, and text+audio. Subtasks include binary fallacy detection and multi-class classification into macro-level fallacy types (Table 1).

In this paper we are targeting fallacy classification and present Team NUST's submission to the shared task. Our key contributions are:

1. We evaluate traditional (SVM, XGBoost) and deep learning models (RoBERTa, Whisper) across uni-modal and multi-modal setups.
2. We propose Prompt-Guided Few-Shot Augmentation (PG-FSA) using GPT-based generation to synthesize fallacy-specific samples for minority classes.
3. We design a late fusion framework combining RoBERTa text and Whisper audio embeddings, enhanced with SMOTE and Focal Loss for better class balance and performance.

We evaluate our framework on the MM-USED-Fallacy dataset under the shared task. Across all three modalities—text-only, audio-only, and text-audio—our method achieved state-of-the-art per-

[*] Equal contribution.
[†] Corresponding author.

[1] MM-ArgFallacy Shared Task

Proceedings of the 12th Argument Mining Workshop, pages 381–387
July 31, 2025 ©2025 Association for Computational Linguistics

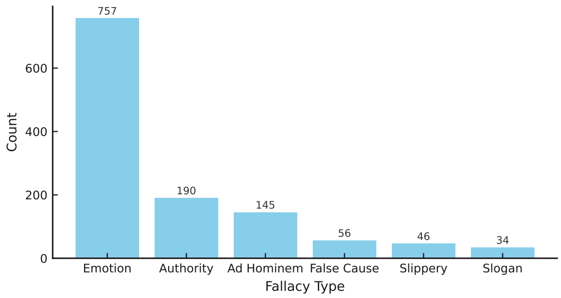

Figure 1: Distribution of fallacy types the training set.

formance, ranking 1st in text-only and multi-modal, and 3rd in audio-only categories. These results validate the effectiveness of our prompt-guided augmentation and modular fusion design.

2 Dataset

The MM-USED-Fallacy dataset, introduced by Mancini et al. (2024b), builds upon textual and audio excerpts from U.S presidential debates. The dataset is obtained using opensource MAMKit tooklkit[2] (Mancini et al., 2024a). Table 1 shows the annotations of dataset into six macro-level fallacy types. Designed for both detection and classification tasks, the dataset supports three modalities: text-only, audio-only, and text+audio.

3 Multi-Class Fallacy Classification

Our proposed framework[3] addresses the dual challenges of data imbalance and modality integration for fallacy classification. It comprises two core components: (1) Prompt-Guided Few-Shot Augmentation (PG-FSA) for data-level augmentation, and (2) Late Fusion Modeling for multi-modal integration. Figure 3 provides an overview of the framework across all modalities.

3.1 Prompt-Guided Few-Shot Augmentation

To mitigate the challenge of class imbalance dataset, we propose Prompt-Guided Few-Shot Augmentation (PG-FSA). This method uses generative capabilities of GPT-4.0, to synthesize high-quality instances for underrepresented fallacy classes. For each minority fallacy category, we engineered a structured prompt which includes formal definition of fallacy from (Goffredo et al., 2022), followed by 15 examples drawn from original training split. Hence, the language model is guided to produce new samples that remain in the semantic boundaries of the target class. To preserve the integrity of the generated samples, all outputs are human-evaluated, the evaluation method and score is discussed in Appendix A.1. This hybrid human-and-model approach allows us to improve minority class representation. The structure of our prompt is given below:

Fallacy Type	Original	PG-FSA	Total
Ad Hominem	145	52	197
False Cause	56	51	107
Slippery Slope	46	50	96
Slogan	34	80	114

Table 2: Sample counts before and after PG-FSA for minority fallacy types.

Prompt

Task:
I want to perform data augmentation because of class imbalance, and this class has very few examples. I want to generate 30 more examples of the class *class_name*.

Class Definition:
definition of class

Instructions:
I have given you 15 examples below from the dataset for your understanding. Study the examples and follow their structuring and other characteristics to generate new examples that align with this definition in the context of the slogans in political debates dataset.

Examples (15 total):
Example 1
Text: *sample from training data 1*
Fallacy Type: *class_name*

Example 2
Text: *sample from training data 2*
Fallacy Type: *class_name*

Figure 2: Prompt for data augmentation in fallacy classification task

Table 2 presents the class-wise augmentation statistics resulting from PG-FSA. We augment the generated samples in the training split given by the organizers. In addition, we also convert these generated textual samples into speech[4] using Eleven Labs[5].

3.2 Methodology

We formulate fallacy classifications as a six-way multi-class classification task spanning three input modalities: text-only, audio-only, and multi-modal (text+audio). The objective is to classify each input instance into one of the six fallacy categories: *Ad Hominem*, *Appeal to Authority*, *Appeal to Emotion*, *False Cause*, *Slippery Slope*, and *Slogan*.

[2] MAMKit Link
[3] Github Link: Source code
[4] By including the synthetic audio clips, the results didn't improve, therefore in the proposed methodology, we employ the orignal data audio clips.
[5] ElevenLabs

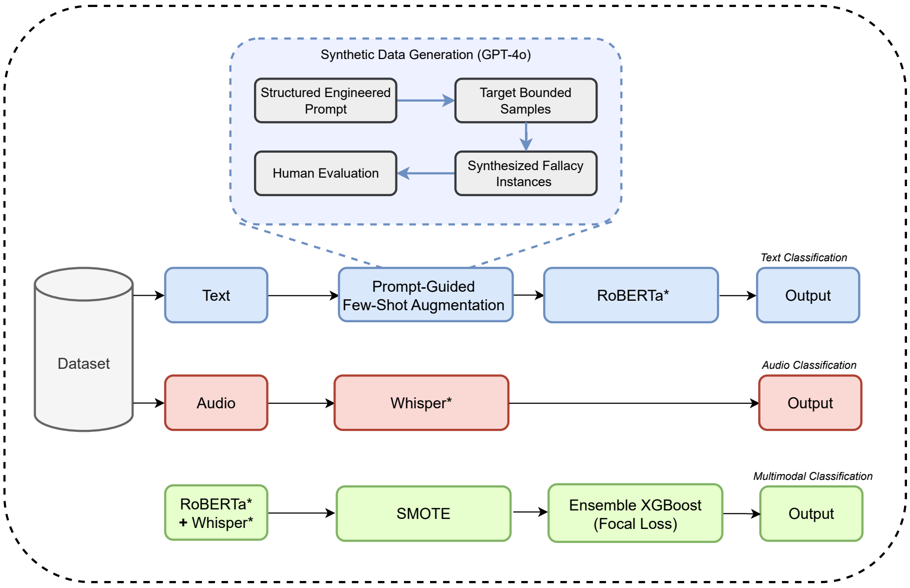
Synthetic Data Generation (GPT-4o)
Structured Engineered Prompt
Target Bounded Samples
Synthesized Fallacy Instances
Human Evaluation
Dataset
Text
Prompt-Guided Few-Shot Augmentation
RoBERTa*
Text Classification
Output
Audio
Whisper*
Audio Classification
Output
RoBERTa* + Whisper*
SMOTE
Ensemble XGBoost (Focal Loss)
Multimodal Classification
Output

Figure 3: Proposed of Multi-Modal Fallacy Classification Framework.

3.2.1 Text-Only Classification

We fine-tune multiple transformer-based language models for text-only fallacy classification, including RoBERTa-small[6] (RoBERTaS), RoBERTa-base[7] (RoBERTaB), RoBERTa-large[8] (RoBERTaL), DeBERTa-base (DeBERTaB), Electra-base (ElectraB), BERT-base[9] (BERTB), and DistilBERT-base[10] (DistilBERTB). Among all models, RoBERTa-base[7] (RoBERTa) model (Liu et al., 2019) with PG-FSA augmentation showed the best performance and is used in the final system. Input utterances are truncated or padded to a maximum sequence length of 128 tokens. A single-layer classification head with six output neurons is appended to the final $[CLS]$ token representation from RoBERTa.

Class Imbalance Mitigation: To address the skewed class distribution, we apply weighted cross-entropy loss. Class weights are inversely proportional to class frequencies, encouraging the model to prioritize minority classes by penalizing their misclassification more heavily. We also experimented with Focal Loss for text only and found it to perform similarly to weighted cross-entropy. To ensure clarity and maintain simplicity in our final presentation, we chose to report only the weighted cross-entropy results.

3.2.2 Audio-Only Classification

We explore both classical and transformer-based pipelines for audio-only classification. For the classical approach, we combine Mel-Frequency Cepstral Coefficients (MFCCs) Feature Extraction (FE) with 2D-CNNs, Gaussian Naive Bayes, Logistic Regression, and SGD Classifier. For transformer-based approach, we fine-tune Whisper (tiny, small, base) (Radford et al., 2022) and Wav2Vec2.0 (Baevski et al., 2020).

We also use PG-FSA augmented data to address class imbalance. We also generate synthetic speech for the augmented textual examples using ElevenLabs[5]' Text to Speech, enabling Whisper and Wav2Vec2.0 to train on both original and synthesized samples. Whisper-small fine-tuned on original data samples gave the best results. We adapt it as an encoder for classification by replacing the decoder with a feedforward layer predicting over six fallacy categories.

All audio inputs are standardized to a sampling rate of 16kHz and fed directly to the Whisper encoder. No text transcriptions are used in this modality.

3.2.3 Multi-modal Fusion

For multi-modal classification, we adopt a late fusion strategy. We encode each modality independently and concatenate them prior to classification. We incorporate RoBERTa-base[7], DistilBERT-base, and their task-specific variants as our text encoders. We use Whisper-small (WhisperS), Whisper with CNN, and Wav2Vec2.0 as our audio encoders. All

[6] smallbenchnlp/roberta-small
[7] FacebookAI/roberta-base
[8] MidhunKanadan/roberta-large-fallacy-classification
[9] mempooltx/bert-base-fallacy-detection
[10] q3fer/distilbert-base-fallacy-classification

the combinations of models used are presented in Table 5.

For fusion of modalities, we first concatenate representations and then we pass the fused representation to a lightweight neural module. It consists of linear projection, layer normalization, ReLU activation and dropout. Final two-layers are feedforward classifier with ReLU activation and dropout regularization. This modular fusion setup enables flexible experimentation with different encoder combinations. Further, we also experiment with various Machine Learning classifiers i.e. Logistic Regression, Random Forest, Gradient Boosting, SVM, and XGBoost+FL. Thus, We evaluate fusion of RoBERTa variants with Whisper, Whisper+CNN, and Wav2Vec2.0 using simple concatenation, XGBoost, and neural projection heads. RoBERTa-base + Whisper-small fused via XGBoost with SMOTE and Focal Loss gave the highest macro-F1 score.

Class Imbalance Mitigation: We adopt two strategies in the multi-modal setting. First, we apply SMOTE (Chawla et al., 2002) in which synthetic samples are generated for minority classes in the fused feature space via interpolation. Second, we use Focal Loss (FL) (Lin et al., 2017) which is used to handle hard-to-classify instances, focal loss down-weights easy examples and focuses the model on minority and ambiguous cases. This dual strategy is chosen to address the increased complexity introduced by the multi-modal setup. The combination of SMOTE and Focal Loss helps balance both underrepresented classes and hard-to-classify examples in the fused feature space.

4 Experimental Setup

This section details the evaluation setup used to benchmark models across three modalities: text-only, audio-only, and multi-modal (text+audio). We organize our discussion into model configurations, fusion strategies, and evaluation metrics.

4.1 Dataset

The official split includes 1,228 training and 2,160 test instances. We use MAMKit loader to obtain the data splits. After applying PG-FSA, the final dataset comprises 1,461 instances. We partition it into training and validation subsets stratified splits. The generated instances are included only in training split.

Model	**M-F1**
Text	
BiLSTM+GloVe	0.4721
RoBERTa	0.3925
RoBERTaB+aug. (Ours)	**0.4856**
Audio	
BiLSTM+MFCCs	0.1582
WavLM	0.0643
WhisperS+aug. (Ours)	**0.1588**
Multi-modal	
BiLSTM-GloVe+MFCCs	0.2191
MM-RoBERTa+WavLM	0.3816
RoBERTaB+WhisperS+XGBoost (Ours)	**0.4611**

Table 3: Macro-F1 scores across modalities. Models marked (Ours) are Team NUST submissions. RoBERTa-base with augmentation (aug.), Whisper-small and RoBERTa-base+Whisper-small+XGBoost performed best.

4.2 Classification Models

We conduct all experiments on a Tesla T4 GPU with 16 GB memory. For text-only models, we use a batch size of 16, max sequence length of 128, and learning rates of $1e-5$ or $2e-5$ depending on model stability. For audio models, the sampling rate is set to 16kHz and maximum audio length is set at 20 seconds, with a batch size of 8. We use AdamW optimizer with early stopping based on validation macro-F1. We use PyTorch and HuggingFace Transformers libraries for all these experiments.

4.3 Evaluation

We use Macro F1 score (M-F1) as the primary evaluation metric due to its robustness in imbalanced multi-class settings. It gives equal importance to each class, making it suitable for assessing performance across both majority and minority fallacy types.

5 Results and Analysis

We evaluate model performance on both validation and official test splits. Table 3 presents the results of the official test set using only the best-performing configurations. Table 3 also presents the baselines are those provided by the shared task organizers. Our models consistently outperform all provided baselines across text-only, audio-only, and multi-modal settings, underscoring the effectiveness of our design choices.

Text-only	M-F1	Audio-only	M-F1
RoBERTaB[7]	0.5441	WhisperS	**0.3168**
RoBERTaL[8]	0.4439	WhisperT	0.1800
DistilBERTB[10]	0.4369	WhisperB w/ FE	0.1275
BERTB[9]	0.3939	Wav2Vec2.0	0.1262
ElectraB	0.3945	Whisper+aug.	0.1260
DeBERTaB	0.4856	Wav2Vec+aug.	0.2400
RoBERTaS	0.4418	MFCC+2D-CNN	0.1281
RoBERTaB (aug. data)	**0.5786**	MFCC+GaussianNB	0.1764
DistilRoBERTaB (aug. data)	0.4418	MFCC+Logistic Regression	0.1622
		MFCC+SGDClassifier	0.1622

Table 4: Macro-F1 scores for various text-only & audio-only models for fallacy classification on the validation set.

Multi-modal Models	M-F1
Pre-Trained	
RoBERTaB[7]+WhisperS	**0.5594**
RoBERTaL[8]+WhisperS	0.5590
DilBERTB[10]+WhisperS	0.4531
RoBERTaB+2D-CNN+Whisper	0.4456
ML Classifiers	
Logistic Regression	0.5438
Random Forest	0.5174
Gradient Boosting	0.5277
SVM	0.5600
XGBoost+FL	**0.5586**

Table 5: Macro-F1 scores for multi-modality models for fallacy classification on the validation set. Fine-tuned neural models and ML classifiers are evaluated using RoBERTa-base and Whisper-based embeddings. Note: RoBERTa and Whisper embeddings are finetuned on MM-Used Fallacy dataset.

5.1 Text-Only

Table 3 shows that our proposed RoBERTa-base[7] model augmented with GPT-generated synthetic data achieves an F1 score of 0.4856 on the test set. This represents a moderate decrease from its validation performance of 0.5786 (reported in Table 4). It is expected given potential variability and distributional differences between the splits. Despite the drop, the model maintains its lead over baselines. This result highlights the benefit of large-scale language models that demonstrate their capabilities through well-structured prompt-driven few-shot generation. Thus, it can mitigate data scarcity and enhance minority class representation.

5.2 Audio-Only

Whisper-small attains a test F1 of 0.1588, down from 0.3168 on the validation split (Appendix, Table 4). While the model slightly outperforms baselines, overall performance remains weak. This suggests that fallacies often lack discriminative acoustic cues, and performance is further degraded by noise, speech clarity issues, and accent variability in the dataset.

5.3 Multi-Modality

Our late fusion model RoBERTa+Whisper with XGBoost achieves F1-score of 0.4611 on the test set (vs. 0.5586 on validation; see Table 5). The model surpasses all baselines, but gains from audio remain limited. Textual features dominate the predictive signal, while simple concatenation may not fully capture cross-modal interactions, particularly for confounding classes like Appeal to Emotion and Slogan. More advanced fusion mechanisms could better align multi-modal features.

5.4 Takeaways

Overall, while all models show some test-time degradation, they consistently outperform baselines. These results emphasize the role of targeted data augmentation and modular design in improving generalization. However, the persistent class imbalance constrains further gains. Future work should focus on advanced augmentation, data cleaning, and robust fusion strategies to unlock better cross-modal alignment and minority class recognition.

6 Conclusions

We tackle the task of fallacy classification across text, audio, and multi-modal inputs under class imbalance constraints. Our framework integrates pretrained models (RoBERTa, Whisper) with prompt-guided few-shot augmentation and late fusion strategies. Experiments on the MM-USED-Fallacy dataset demonstrate strong validation and test performance across all modalities. RoBERTa-base[7] with augmentation proves most effective for text, Whisper-small performs best for audio, and late fusion with XGBoost yields the highest multi-modal gains. Future directions include modality alignment, adaptive fusion, and contrastive learning to enhance cross-modal reasoning and representation.

Limitations

While our framework achieves strong performance across modalities, a few limitations remain:

Simple Fusion Strategy: We adopt a late fusion approach using feature concatenation followed by XGBoost. While effective, this strategy may not fully capture fine-grained inter-modal dependencies. More advanced fusion techniques (e.g., cross-attention or tensor fusion) could potentially yield better alignment between modalities.

Limited Use of Context: Although contextual utterances are provided in the dataset, our current setup does not explicitly model discourse-level dependencies. Incorporating contextual reasoning (e.g., via hierarchical transformers or dialogue-aware models) may improve understanding of fallacies with pragmatic cues.

Synthetic Data Quality: Prompt-guided augmentation boosts performance, especially for underrepresented classes, but generated samples may vary in linguistic quality or realism. Filtering or scoring mechanisms could help ensure higher fidelity in future iterations.

Underperformance in Audio Modality: Despite outperforming baselines, audio-only models remain weaker due to the inherently low signal-to-noise ratio in acoustic fallacy cues. Improvements could be made via better preprocessing (*e.g.*, noise suppression, speaker normalization) or pretrained models fine-tuned for prosodic features.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. In *Journal of Artificial Intelligence Research*, volume 16, pages 321–357.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based Detection and Classification of Fallacies in Political Debates. In *ACL Anthology*, volume 2023.findings-emnlp.684 of *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11101–11112, Singapore (SG), Singapore. 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-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 69–82, Bangkok, Thailand. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 170–178, St. Julian's, Malta. Association for Computational Linguistics.

Alec Radford, Jong Wook Gao, Greg Brockman, Vicki Narasimhan, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, and 1 others. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

A Appendix

A.1 Human Evaluation

To ensure the quality of the augmented examples, we use a two-stage human evaluation process. Two independent annotators assess each example against a predefined evaluation criterion to determine whether it matches the intended class. Examples with mutual agreement on label 1 are retained, while those with agreement on label 0 are discarded. In cases of disagreement, the annotators conduct a follow-up discussion to reach a consensus, and the agreed label is marked as the final evaluation. The final augmented dataset includes only examples with a final label of 1. An inter-annotator agreement, measured as raw percentage agreement (due to the absence of negative examples), is 87.55%.