# Overview of MM-ArgFallacy2025 on Multimodal Argumentative Fallacy Detection and Classification in Political Debates

**Eleonora Mancini[1*], Federico Ruggeri[1], Serena Villata[2], Paolo Torroni[1]**

[1]DISI, University of Bologna
{e.mancini, federico.ruggeri6, p.torroni}@unibo.it
[2]Université Côte d'Azur, Inria
serena.villata@inria.fr

## Abstract

We present an overview of the MM-ArgFallacy2025 shared task on Multimodal Argumentative Fallacy Detection and Classification in Political Debates, co-located with the 12th Workshop on Argument Mining at ACL 2025. The task focuses on identifying and classifying argumentative fallacies across three input modes: text-only, audio-only, and multimodal (text+audio), offering both binary detection (AFD) and multi-class classification (AFC) subtasks. The dataset comprises 18,925 instances for AFD and 3,388 instances for AFC, from the MM-USED-Fallacy corpus on U.S. presidential debates, annotated for six fallacy types: Ad Hominem, Appeal to Authority, Appeal to Emotion, False Cause, Slippery Slope, and Slogan. A total of 5 teams participated: 3 on classification and 2 on detection. Participants employed transformer-based models, particularly RoBERTa variants, with strategies including prompt-guided data augmentation, context integration, specialised loss functions, and various fusion techniques. Audio processing ranged from MFCC features to state-of-the-art speech models. Results demonstrated textual modality dominance, with best text-only performance reaching 0.4856 F1-score for classification and 0.34 for detection. Audio-only approaches underperformed relative to text but showed improvements over previous work, while multimodal fusion showed limited improvements. This task establishes important baselines for multimodal fallacy analysis in political discourse, contributing to computational argumentation and misinformation detection capabilities.

## 1 Introduction

This paper presents an overview of MM-ArgFallacy2025: multimodal argumentative fallacy detection and classification on political debates; the task is organized for the first time.

In the past decade, several studies have highlighted the importance of Argument Mining on semantic textual analysis, leading to a broad set of applications, including legal analytics, social media, and biomedicine, to name a few. However, past research has also theorized the importance of including paralinguistic features in argumentative discourse analysis to capture additional dynamics that cannot be extracted from text alone. Consequently, Multimodal Argument Mining (MAM) emerged, aiming to validate these propositions empirically and gain a more comprehensive understanding of argumentative discourse by integrating multiple modalities. So far, core argument mining tasks like argument detection, component classification, and relation classification have been mainly explored, where the integration of audio modality has proved to be effective. Recently, other tasks like fallacy detection and classification have been investigated in the context of MAM, but they are still underexplored.

MM-ArgFallacy2025[1] aims to advance research in this latter area by providing a platform for the development and the evaluation of systems capable of detecting and classifying argumentative fallacies using different modalities. Specifically, MM-ArgFallacy2025 challenges participants to distinguish whether a given sentence from a political debate contains an argumentative fallacy and, if any, which type of logical inconsistency is observed. This research is crucial for advancing NLP technologies and accelerating adoption for user benefit, contributing to the development of systems for aiding users in knowledge acquisition and awareness of controversial topics, consistent with AM contributions in social media (Dusmanu et al., 2017; Lytos et al., 2019) and debates (Carstens et al., 2014; Swanson et al., 2015; Haddadan et al., 2019).

[1] https://nlp-unibo.github.io/mm-argfallacy/2025/



*Proceedings of the 12th Argument Mining Workshop*, pages 358–368

With the integration of audio modality, we enrich the spectrum of available features for studying human fallacies, fostering the development of more accurate models. We build on existing work and focus on the political debates, where argumentative content and reasoning fallacies are abundant. The task concerns two sub-tasks: detecting argumentative fallacious sentences and classifying them. We follow (Mancini et al., 2022) and consider three input modes in each sub-task for assessing individual modalities in addition to the multimodal setting: **text-only**, where only an input textual sentence from a political debate dialogue is provided; **audio-only**, where only an audio sample corresponding to a textual sentence from a political debate dialogue is provided; **text-audio**, where an input textual sentence from a political debate dialogue and its aligned audio sample are provided.

We evaluate the participating systems based on binary F1 score for fallacy detection and macro-averaged F1 score for fallacy classification. The latter metric balances precision and recall across fallacious categories, ensuring a fair assessment of system performance. The task's comprehensive evaluation framework, coupled with the diverse multimodal datasets, provides a rigorous benchmark for advancing the state-of-the-art in MAM and, in particular, in multimodal fallacy recognition.

In the rest of the paper, we offer an overview of MM-ArgFallacy2025, detailing the datasets, the evaluation measures, and the submission guidelines. We also present the results and the methodologies of the participating systems, highlighting the progress and the challenges when developing robust MAM solutions. By fostering collaboration and innovation in this critical area, MM-ArgFallacy2025 contributes to the broader goal of enhancing the reliability of automated content analysis in the digital age.

## 2 Related Work

The study of fallacies is deeply rooted in argumentative theory, which dates back to Aristotle [2]. The utility of recognizing and studying fallacies mainly emerged in the '70-80s (Hamblin, 2022). Moreover, fallacy detection is directly related to human reasoning, where communication can often degenerate into conflicts, disagreements, and debates due to logical fallacies in discourse (Jin et al., 2022). Detecting and classifying fallacies in discourse is a valuable tool in several applications, including analyzing human behavior in dialogical settings, preventing misinformation spread in fact-checking systems, and evaluating generative models' reasoning. Research on fallacy detection and classification is not limited to text analysis, but could encompass other modalities too, such as audio, where paralinguistic features can often be associated with specific fallacy types (Kišiček, 2020a).

Research on the interplay between arguments and emotions in speech began with Benlamine et al. (Benlamine et al., 2015). Subsequently, further studies focused on multimodality in argumentation, showing the correlation of paralinguistic features with argumentative discourse in various domains, including advertisements, news coverage, and legal analytics (Kišiček, 2014; Groarke and Kišiček, 2018; Kišiček, 2020b). These findings led to the development of Multimodal Argument Mining (MAM), where Lippi and Torroni (2016) conducted the first study in political debates, focusing on UK ministerial elections but limited to a single debate.

Interest in political debates motivated further research for argumentative tasks like argument component detection and classification (Haddadan et al., 2019), argumentative fallacy classification (Goffredo et al., 2022a, 2023), and argumentative relation identification (Mestre et al., 2023). Recent work has particularly focused on multimodal approaches to these argumentative tasks in political contexts (Mancini et al., 2022, 2024b; Mestre et al., 2023).

## 3 Problem Formulation

MM-ArgFallacy2025's subtasks are formulated as follows.

**Argumentative Fallacy Detection (AFD).** The input is a sentence, in the form of text or audio or both, extracted from a political debate. The objective is to determine whether the input contains an argumentative fallacy.

**Argumentative Fallacy Classification (AFC).** The input is a sentence, in the form of text or audio or both, extracted from a political debate, containing a fallacy. The objective is to determine the type of fallacy contained in the input, according to the classification introduced by (Goffredo

[2] https://plato.stanford.edu/entries/fallacies/

| Snippet | Fallacy Category |
|---|---|
| *the same kind of woolly thinking* | Appeal to Emotion |
| *As George Will said the other day, "Freedom on the march; not in Russia right now."* | Appeal to Authority |
| *Governor Carter apparently doesn't know the facts.* | Ad Hominem |
| *We won the Cold War because we invested and we went forward.* | False Cause |
| *And if we don't act today, the problem will be valued in the trillions.* | Slippery Slope |
| *We have to practice what we preach.* | Slogan |

Table 1: Examples of annotated fallacies.

et al., 2022a).[3] In particular, the fallacy categories are: Appeal to Emotion, Appeal to Authority, Ad Hominem, False Cause, Slippery Slope, and Slogan. Table 1 reports examples of each fallacy category.

For each sub-task, participants can leverage the debate context of a given input: all its previous sentences and corresponding aligned audio samples. For instance, consider the **text-only** input setting. Given a sentence from a political debate at index $i$, participants can use sentences with indexes from 0 to $i-1$, where 0 denotes the first sentence in the debate.

## 4 Data

We describe the available training data for the challenge and the data collection process to curate the test set used to evaluate participants' submissions in the challenge (hereinafter, denoted as **secret test set**). All datasets are made available through MAMKit (Mancini et al., 2024a)[4]. Since most of these multimodal datasets cannot release audio samples for copyright reasons, MAMKit provides a simple interface to dynamically build them and foster reproducible research.

### 4.1 Training Data

The primary training dataset is **MM-USED-fallacy** (Mancini et al., 2024b). The data provides annotations for AFC and AFD subtasks. The dataset comprises 1,228 fallacies with corresponding context information from the dataset of (Haddadan et al., 2019) on US presidential elections. The fallacies are labeled as argumentative fallacies belonging to six categories introduced in (Goffredo et al., 2022a). Additionally, inspired by (Goffredo et al., 2022a)'s observations on the benefits of employing other argument mining tasks like component detection for fallacy detection and classification, participants could use the following datasets to encourage multi-task training approaches (see Table 2 for a summary).

**UKDebates (Lippi and Torroni, 2016).** A dataset of 386 sentences and corresponding audio samples about three candidates for the UK Prime Ministerial elections of 2015. Sentences are annotated for argumentative sentence detection: a sentence is labeled as containing or not containing a claim.

**M-Arg (Mestre et al., 2021a).** A multimodal dataset built around the 2020 US Presidential elections for argumentative relation classification: a sentence can attack, support, or have no relation with another sentence. The dataset contains 4,104 sentence pairs and corresponding audio sequences of four candidates and a debate moderator concerning 18 topics. A high-quality subset of M-Arg is also provided, containing 2,443 sentence pairs with high agreement confidence.

**MM-USED (Mancini et al., 2022).** A multimodal extension of the USElecDeb60to16 dataset introduced in (Haddadan et al., 2019). It contains presidential candidates' debate transcripts and corresponding audio recordings aired from 1960 to 2016. The dataset contains 26,781 labeled sentences and corresponding audio samples from 39 debates and 26 distinct speakers for argumentative sentence detection and argumentative component classification: a sentence can contain a claim, a premise, or neither of them.

**Text to Audio Alignment Corrections.** Compared to the initial release of MAMKit (Mancini et al., 2024a), we introduce an improved text-to-audio alignment framework based on WhisperX (Bain et al., 2023), a state-of-the-art speech recognition model that allows for precise and fine-grained audio-to-text alignment. We use this framework to address some well-known alignment issues in MM-USED (Mancini et al., 2022) and MM-USED-fallacy (Mancini et al., 2024b), allowing to integrate previously discarded debates and favouring the collecting of novel data (4.2). The updated datasets are available through MAMKit.

[3] We only refer to macro categories while sub-categories are left for future work.

[4] https://nlp-unibo.github.io/mamkit/

| Name | No. Samples | Task$^{\alpha}$ | Domain |
|---|---|---|---|
| **Primary** | | | |
| MM-USED-fallacy | 18,925; 3,338 | AFD; AFC | US Presidential Elections |
| **Supplementary** | | | |
| UKDebates | 386 | ASD | UK Prime Ministerial Elections |
| M-ARG | 4,104 / 2,443 | ARC | US Presidential Elections |
| MM-USED | 26,781 | ASD, ACC | US Presidential Elections |

Table 2: Available datasets for MM-ArgFallacy2025 shared task.
$^{\alpha}$Following (Mancini et al., 2024a), we denote tasks as Argumentative Fallacy Detection (AFD), Argumentative Fallacy Classification (AFC), Argumentative Sentence Detection (ASD), Argumentative Relation Classification (ARC), Argumentative Component Classification (ACC).

### 4.2 Test Data

**Data Collection.** We collect and annotate novel debates from US politics available in The American Presidency Project[5]. In particular, we consider the first presidential debate of the election cycle in Atlanta between Trump and Biden, aired on 28th June 2024, and the first presidential debate between Trump and Harris, aired on 11 September 2024. We follow the data collection pipeline proposed by Mancini et al. (2022) to retrieve original audio recordings, but improve the text-to-audio alignment by leveraging WhisperX (Bain et al., 2023) for transcription, alignment and diarization. We follow Goffredo et al. (2022a) and split debates into paragraphs, where each paragraph corresponds to a speaker turn. We use our text-to-audio alignment framework to pair paragraphs and corresponding textual sentences with related audio chunks. In total, we obtain 134 paragraphs for the first debate and 163 for the second one.

**Data Annotation.** For the annotation, we instruct two annotators with expertise in AM tasks and near-to-native English proficiency. We provide annotators with the guidelines of Goffredo et al. (2022a) for detecting and classifying argumentative fallacies. This is required to ensure annotation consistency with existing datasets (e.g., MM-USED-fallacy). We remove paragraphs (i.e., dialogue turns) that do not belong to the main speakers of the debate (i.e., Trump, Biden and Harris). In total, we obtain 154 paragraphs for annotation, equivalent to 2154 sentences. We rely on Label Studio[6] for annotation, an open-source data annotation platform. For AFD and AFC subtasks, we provide annotators with the same instructions described in Section 3. In particular, for AFD, annotators label each sentence in a given paragraph from a debate as containing a fallacy. In case an annotator labels a sentence as fallacious, they also provide the corresponding logical fallacy category to address AFC subtask.

**Inter-Annotator Agreement.** Since fallacies can span multiple sentences (Goffredo et al., 2022a), we report the rate of exact and partial overlaps between annotations. An exact overlap is when both annotators agree on all sentences constituting a fallacy. In contrast, a partial overlap is when annotators agree on a subset of sentences constituting a fallacy. We observe 236 overlaps, 110 of which are exact, while the remaining 126 are partial. The agreement rate measured as the number of sentences detected as fallacious of the same category by both annotators is 67.37%. Moreover, we compute inter-annotator agreement (IAA) at sentence level, measured as Cohen's Kappa (Cohen, 1968). For AFD, the IAA is 0.4787 (*moderate agreement*), while, for AFC, the IAA is 0.4954 (*moderate agreement*). Additionally, regarding AFC, the per category IAA is as follows: 0.411 Appeal to Emotion, 0.337 Appeal to Authority, 0.357 Ad Hominem, 0.224 False Cause, and 0.712 Slogan. No annotator labeled a fallacy instance as Slippery Slope.

**Resulting Dataset.** Table 3 reports the statistics of the resulting secret test set. We observe that a large majority of fallacies belong to Appeal to Emotion, followed by Ad Hominem. These findings are in line with annotations reported in previous work on argumentative fallacy classification (Goffredo et al., 2022a). Regarding AFD, the collected secret test set contains 229 fallacies and 1,946 non-fallacious sentences.

[5] https://www.presidency.ucsb.edu/
[6] https://labelstud.io/

| Fallacy Category | No. Instances |
|---|---|
| Appeal to Emotion | 142 |
| Appeal to Authority | 16 |
| Ad Hominem | 46 |
| False Cause | 16 |
| Slippery Slope | 0 |
| Slogan | 9 |
| Total | 229 |

Table 3: Secret test set statistics for Argumentative Fallacy Classification.

## 5 Overview of the Systems and Results

Three teams participated in the AFC subtask, while two teams participated in the AFD subtask. In total, participants submitted 25 valid runs. No team participated in both subtasks.

Table 4 shows the results achieved by the individual teams for each subtask. Regarding AFC, we observe that only one team, Team NUST (Tahir et al., 2025), beats baselines on the text-only modality with a F1-score of 0.4856. This result shows that even a simple baseline like a BiLSTM is a strong competitor. In contrast, all participants improved over the baselines when considering the audio modality, while two teams surpassed the transformer baseline in the multimodal setting. Regarding AFD, Team Ambali_Yashovardhan reaches comparable performance to baselines in the text-only input setting, achieving rank 2. Nonetheless, despite reporting significant results in the audio-only setting where baselines fail the task, their solution is outperformed by both baselines in the multimodal setting.

All teams used neural networks, with transformer-based models being the most frequent choice. Some teams also employed machine learning classifiers like XGBoost on top of neural network models. Moreover, several teams explored a wide set of solutions to account for class imbalance, a well-known challenge in fallacy detection (Goffredo et al., 2022a).

### 5.1 Baselines

For both tasks, we employ the same set of baselines: a feature-based BiLSTM (Mancini et al., 2024b) and a transformer-based model (Mancini et al., 2024a). Regarding the BiLSTM model, the baseline uses GloVe embeddings for text inputs and extracted MFCCs features for audio recordings. Conversely, the transformer-based model uses RoBERTa (Liu et al., 2019)[7] for encoding text and WavLM (Chen et al., 2022)[8] for audio. Both architectures employ a late fusion strategy for the multimodal setting, where text and audio embeddings are concatenated and fed to a final classification layer. Independently of the given input setting, we denote the baselines as **Baseline BiLSTM** and **Baseline Transformer**, respectively.

### 5.2 System Descriptions and Task-Specific Results

Below, we describe the approaches of all participating systems; see also Table 5 for an overview.

#### 5.2.1 Argumentative Fallacy Classification.

Team **NUST** (Tahir et al., 2025) employ RoBERTa (Liu et al., 2019) for text encoding and Whisper (Radford et al., 2023) for audio encoding. The two encodings are combined in a late fusion fashion without requiring joint end-to-end training and fed to a XGBoost (Chen and Guestrin, 2016) classifier. To account for label imbalance, they propose several solutions, including generating synthetic samples via GPT 4.0, class weighting, SMOTE (Chawla et al., 2002) in which synthetic samples are generated for minority classes in the fused feature space via interpolation, and focal loss (Lin et al., 2017) to handle hard-to-classify instances.

Team **AlessioPittiglio** (Pittiglio, 2025) explore a wide set of transformer-based text and audio encoders. For text, they evaluate BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2023) and ModernBERT (Warner et al., 2024). In particular, the authors propose three different strategies to integrate context information by (i) concatenating tokens during tokenization, (ii) concatenating pooled embeddings, and (iii) leveraging cross-attention. For audio, they evaluate Wa2Vec 2.0 (Baevski et al., 2020), WavLM (Chen et al., 2022) and HuBERT (Hsu et al., 2021). Regarding multimodality, they build an ensemble of the best text (RoBERTa with context information) and audio (HuBERT) models via a weighted average of individual encoder logits. Weights are calibrated via Bayesian optimization (Snoek et al., 2012).

Team **CASS** (Kalyan et al., 2025) encodes texts with RoBERTa (Liu et al., 2019) and audio

[7] *FacebookAI/roberta-base*
[8] *patrickvonplaten/wavlm-libri-clean-100h-base-plus*

| Team | Rank | Text | Rank | Audio | Rank | Text-Audio |
|---|---|---|---|---|---|---|
| | | | AFC | | | |
| Team NUST | 1 | **0.4856** | 2 | 0.1588 | 1 | **0.4611** |
| Team AlessioPittiglio | 3 | 0.4444 | 1 | **0.3559** | 2 | 0.4403 |
| Team CASS | 5 | 0.1432 | 4 | 0.0864 | 5 | 0.1432 |
| Baseline BiLSTM | 2 | 0.4721 | 3 | 0.1582 | 4 | 0.2191 |
| Baseline Transformer | 4 | 0.3925 | 5 | 0.0643 | 3 | 0.3816 |
| | | | AFD | | | |
| Team Ambali_Yashovardhan | 2 | 0.2534 | 1 | **0.2095** | 3 | 0.2244 |
| Team EvaAdriana | 4 | 0.2195 | 2 | 0.1690 | 4 | 0.1931 |
| Baseline BiLSTM | 3 | 0.2462 | 3 | 0.0000 | 2 | 0.2337 |
| Baseline Transformer | 1 | **0.2770** | 3 | 0.0000 | 1 | **0.2848** |

Table 4: Results for multimodal argumentative fallacy detection on political debates. For AFC, we report the macro F1-score, while for AFD, we report the binary F1-score. Best results per subtask are in **bold**, second best results are underlined.

| Team | Task | | Text | | | | | | | Audio | | | | | MAM | | | Misc | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AFC | AFD | BERT | RoBERTa | DeBERTa | ModernBERT | SBERT | ALBERT | DeepSeek R1 | Whisper | Wav2Vec | WavLM | HuBERT | MFCCs | Late Fusion | Early Fusion | Mid Fusion | Data Augmentation | Class Weighting | Bayesian Optimization | Focal Loss | Context |
| NUST (Tahir et al., 2025) | ✓ | | | ✓ | | | | | | ✓ | | | | | ✓ | | | ✓ | ✓ | | | |
| AlessioPittiglio (Pittiglio, 2025) | ✓ | | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| CASS (Kalyan et al., 2025) | ✓ | | | ✓ | | | | | | | ✓ | | | ✓ | ✓ | | | | | | | |
| Ambali_Yashovardhan | | ✓ | | ✓ | | | | | | | | | ✓ | | ✓ | | | | | | ✓ | |
| EvaAdriana (Larumbe and Vendrell, 2025) | | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ | | ✓ | | | ✓ | ✓ | | | | | | | |

Table 5: **Overview of the approaches.** The numbers in the language box refer to the position of the team in the official ranking.

recordings using a BiLSTM with extracted Mel-Frequency Cepstral Coefficients (MFCCs) features as input (Mancini et al., 2024b). For multimodality, they concatenate the pooled embeddings of RoBERTa for text and Wav2Vec (Baevski et al., 2020) for audio. The concatenated embedding is eventually fed to a logistic regression classifier.

### 5.3 Argumentative Fallacy Detection

Team **Ambali_Yashovardhan** uses RoBERTa (Liu et al., 2019) for processing text inputs and Distil-HuBERT (Chang et al., 2022) for encoding audio recordings. Due to memory constraints, they limit audio sequence length to 320,000 samples (approximately 20 seconds at 16 kHz), truncating longer files. Regarding multimodality, the authors adopt a late fusion strategy where a weighted average of each modality model's logits is computed. The weights are learnt during training. To handle class imbalance, they adopt focal loss (Lin et al., 2017).

Team **EvaAdriana** (Larumbe and Vendrell, 2025) explore five transformer-based models for text modality: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), SBERT (Reimers and Gurevych, 2019), ALBERT (Lan et al., 2020), and DeepSeek R1 (DeepSeek-AI et al., 2025). Up to the last four layers of encoder-based transformers are unfrozen for fine-tuning on the task, while DeepSeek is 4-bit quantized to account for the available computational budget. Regarding audio modality, the authors evaluate two models: a CNN with MFCCs features as inputs and Wav2Vec 2.0 (Baevski et al., 2020). Lastly, they combine the best-performing models for text (RoBERTa) and audio (Wa2Vec 2.0) for multimodality. In particular, they concatenate the output of each modality encoder and feed it to a classification head.

## 6 Main Findings

The MM-ArgFallacy2025 shared task revealed key insights about multimodal fallacy analysis through binary detection and multi-class classification tasks, highlighting both capabilities and fundamental limitations in this field.

**Textual Dominance.** Text-based approaches consistently outperformed audio and multimodal alternatives across all teams in AFC. Team NUST achieved a 0.4856 F1-score for classification, establishing transformer-based models, particularly RoBERTa variants, as the most effective. Team EvaAdriana's comparative evaluation in AFD revealed that fine-tuned transformer models substantially outperformed zero-shot approaches, with their task-specific RoBERTa achieving 0.3393 F1-score compared to DeepSeek-R1's zero-shot performance of 0.1567 on the validation set, underscoring the importance of domain-specific adaptation for specialised argumentation tasks.

**Audio Processing Challenges.** Audio-only approaches consistently underperformed, with Team AlessioPittiglio's best performance of 0.3559 remaining substantially below text baselines despite employing state-of-the-art models including HuBERT, Wav2Vec2, and Whisper. However, the results achieved by the participants demonstrate progress over previous work in extracting fallacy-relevant acoustic features. Technical constraints forced audio truncation to 15-20 seconds, but more fundamentally, acoustic signatures of fallacious reasoning appear too subtle for current speech processing models to reliably capture.

**Limited Multimodal Gains.** Multimodal approaches failed to deliver expected performance improvements. Team NUST's late fusion achieved 0.4611, only modestly improving over text baselines while requiring significant computational cost. This suggests simple fusion strategies are insufficient to capture complex relationships between semantic content and paralinguistic delivery, with textual information overwhelming rather than complementing audio features.

**Effective Strategies.** Team NUST's success stemmed from prompt-guided augmentation using GPT-4. Team AlessioPittiglio's context integration (3-4 previous sentences) contributed to strong performance, though benefits were primarily textual, suggesting modality-specific strategies are needed.

**Class Imbalance Challenge.** Severe class imbalance emerged as the primary technical challenge. Binary detection faced 90.8% vs. 9.2% distribution, while classification presented "double imbalance" with *Appeal to Emotion* comprising 59% of fallacious samples versus <3% for minority classes. Team NUST's synthetic data generation proved more effective than algorithmic adjustments, indicating quality augmentation outperforms technical modifications for addressing imbalance.

## 7 Discussion

The MM-ArgFallacy2025 shared task established important baselines while revealing both progress and limitations in multimodal fallacy analysis.

**Audio Improvements.** While text-based methods achieved the highest performance across both detection and classification tasks, teams demonstrated notable improvements in audio-only approaches compared to previous work, with Team AlessioPittiglio achieving a 0.3559 F1-score for classification using HuBERT-based models. These advances suggest that audio modalities contain valuable information for fallacy detection, though current extraction techniques remain limited.

**Speaker Dependency.** The challenges in audio-only approaches may also be attributed to the inherently speaker-dependent nature of acoustic cues. As noted by previous work, different speakers have varying skills in using vocal cues such as articulation, sonority, and tempo, and possess different levels of persuasive power, with vocal characteristics directly affecting the clarity, credibility, and receptivity of a speaker's message (Lippi and Torroni, 2016).

**Multimodality Fusion Strategies.** The reliance on simple concatenation and late fusion approaches in the proposed multimodal systems reveals fundamental gaps in current methodologies. These approaches fail to capture complex interdependencies between linguistic content and paralinguistic delivery, which likely explains the limited performance of many multimodal systems. Rather than indicating the non-effectiveness of multimodal integration itself, these results highlight the need for more sophisticated fusion architectures that can jointly learn complementary cues from both modalities during training. Future work should prioritise advanced fusion architectures that enable joint

learning across modalities, moving beyond late fusion toward cross-attention mechanisms and early integration strategies.

**Conditional Audio Generation.** Key contributions of this shared task include successful prompt-guided data augmentation and systematic context integration strategies that proved effective for addressing severe class imbalance and improving classification performance. Building on the success of synthetic textual data generation, a possible direction involves controllable conditional audio generation that transcends the basic text-to-speech approaches employed by teams in this shared task. Drawing inspiration from recent advances in signal-to-language augmentation (Kumar et al., 2024) that enable fine-grained control over acoustic parameters such as loudness, pitch, reverb, brightness, and duration, future research could develop fallacy-aware audio generation systems. Such approaches go beyond traditional digital signal processing by incorporating learned representations that capture how acoustic characteristics convey persuasive intent and logical flaws in context. For appeals to emotion, generation could emphasize particular intonation patterns and vocal intensity, while deceptive reasoning patterns could incorporate vocal stress indicators, hesitation markers, or pitch variations suggesting uncertainty.

## 8 Conclusion

The MM-ArgFallacy2025 shared task demonstrates that fallacy detection and classification remain challenging problems with significant potential for advancement. Text-based approaches currently show the most promise, while audio and multimodal systems require architectural innovations to realise their full potential. The ultimate goal remains developing integrated systems that effectively leverage both semantic and paralinguistic cues to support democratic discourse and critical thinking education.

## Limitations

**Annotations.** In alignment with (Mancini et al., 2024b), we advocate for approaching fallacy classification as a multimodal problem. Nonetheless, the annotation methodology employed in this study mirrors that of (Goffredo et al., 2022b), relying solely on textual information for both training data preparation and secret test set creation. This text-centric approach potentially overlooks crucial information embedded in the acoustic characteristics of spoken debates, such as intonational patterns, emphasis, and other paralinguistic features that could indicate fallacious arguments. Achieving the full potential of multimodal fallacy detection will require developing new annotation protocols that systematically integrate both linguistic and acoustic dimensions from the ground up.

**MAMKit.** MAMKit remains an evolving toolkit with several acknowledged limitations that reflect its ongoing development status. The platform currently supports only PyTorch, which may present integration challenges for researchers working with alternative frameworks or seeking to incorporate existing work built on different architectures. Additionally, the toolkit's coverage of multimodal argumentation resources is non exhaustive, as several established datasets (e.g., VivesDebate-Speech (Ruiz-Dolz and Iranzo-Sánchez, 2023), ImageArg (Liu et al., 2022), MMClaims (Cheema et al., 2022)) and models (e.g., M-ArgNet (Mestre et al., 2021b)) have not yet been integrated. Furthermore, the current scope is restricted to text and audio modalities, excluding visual argumentation mining despite its growing importance in the field. Nevertheless, deploying MAMKit to deliver the datasets used in this shared task provided valuable opportunities to gather community feedback and identify priority areas for future development, informing our roadmap for expanding both framework compatibility and multimodal coverage.

**Dataset Scale and Imbalance.** The MM-USED-fallacy dataset faces dual constraints that significantly impact model development and evaluation. First, with only 3,388 instances for the AFC task, the dataset represents a relatively small scale for training robust deep learning models, a limitation characteristic of specialised argumentation tasks where high-quality annotations are resource-intensive to obtain. The expansion conducted for this shared task, while methodologically sound, added only 229 fallacious instances from two 2024 debates, maintaining the dataset's modest scale. Second, severe class imbalance permeates both detection and classification tasks, with fallacious sentences comprising merely 9.2% of instances in binary detection, while classification exhibits "double imbalance" with Appeal to Emotion representing 59% of fallacious samples versus minority classes accounting for less than 3% each. Some fallacy types, such as Slippery Slope, are com-

pletely absent from the test set, preventing comprehensive evaluation. These scale and distribution constraints compound each other, limiting model generalization capabilities across diverse speaking styles, debate formats, and political contexts while making robust performance assessment particularly challenging for underrepresented fallacy categories. Future work should prioritize systematic dataset expansion across multiple election cycles and speaker demographics while developing targeted annotation strategies to achieve more balanced fallacy type distributions.

## Acknowledgments

This work was partially supported by project "FAIR - Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", under the European Commission's NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, Partenariato Esteso (PE00000013). F. Ruggeri is partially supported by the European Union's Justice Programme under Grant Agreement No. 101087342 for the project "Principles Of Law In National and European VAT".

## References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech 2023*, pages 4489–4493.

M. Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. Emotions in argumentation: an empirical evaluation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 156–163. AAAI Press.

Lucas Carstens, Francesca Toni, and Valentinos Evripidou. 2014. Argument mining and social debates. In *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 451–452. IOS Press.

Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. 2022. Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7087–7091.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

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.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022a. 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.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022b. 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.

Leo Groarke and Gabrijela Kišiček. 2018. Sound arguments: An introduction to auditory argument. In *Argumentation and inference: Proceedings of 2nd European Conference on Argumentation*, pages 177–198. London: Collage Publications.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Charles Leonard Hamblin. 2022. Fallacies. In *Advanced Reasoning Forum*. Advanced Reasoning Forum.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Warale Avinash Kalyan, Siddharth Pagaria, Chaitra V, and Spoorthi H G. 2025. Multimodal argumentative fallacy classification in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*.

Gabrijela Kišiček. 2014. The role of prosodic features in the analysis of multimodal argumentation. In *International Society for the Study of Argumentation (ISSA), 8th international conference on argumentation*. Rozenberg Quarterly, The Magazine.

Gabrijela Kišiček. 2020a. Listen carefully! fallacious auditory arguments. In *Proceedings of the 12th Conference of the Ontario Society for the Study of Argumentation*, OSSA 12, pages 17–32. University of Windsor.

Gabrijela Kišiček. 2020b. Listen carefully! fallacious auditory arguments. In *Proceedings of the 12th Conference of the Ontario Society for the Study of Argumentation*, OSSA 12, pages 17–32. University of Windsor.

Sonal Kumar, Prem Seetharaman, Justin Salamon, Dinesh Manocha, and Oriol Nieto. 2024. Sila: Signal-to-language augmentation for enhanced control in text-to-audio generation. *Preprint*, arXiv:2412.09789.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Eva Cantín Larumbe and Adriana Chust Vendrell. 2025. Argumentative fallacy detection in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2979–2985. AAAI Press.

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. *Preprint*, arXiv:1907.11692.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055.

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, Andrea Galassi, and Paolo Torroni. 2022. Multimodal argument mining: A case study in political debates. In *Proceedings of the 9th Workshop on Argument Mining*, pages 158–170, Online and in Gyeongju, Republic of Korea. International Conference on 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.

Rafael Mestre, Stuart E. Middleton, Matt Ryan, Masood Gheasi, Timothy Norman, and Jiatong Zhu. 2023. Augmenting pre-trained language models with audio feature embedding for argumentation mining in political debates. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 274–288, Dubrovnik, Croatia. Association for Computational Linguistics.

Rafael Mestre, Razvan Milicin, Stuart Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021a. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *ArgMining@EMNLP*, pages 78–88. Association for Computational Linguistics.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021b. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessio Pittiglio. 2025. Leveraging context for multimodal fallacy classification in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ramon Ruiz-Dolz and Javier Iranzo-Sánchez. 2023. VivesDebate-speech: A corpus of spoken argumentation to leverage audio features for argument mining. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2071–2077, Singapore. Association for Computational Linguistics.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, page 2951–2959, Red Hook, NY, USA. Curran Associates Inc.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Abdullah Tahir, Imaan Ibrar, Huma Ameer, Mehwish Fatima, and Seemab Latif. 2025. Prompt-guided augmentation and multi-modal fusion for argumentative fallacy classification in political debates. *"Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025)"*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.