





Multimodal Fallacy Classification in Political Debates

Eleonora Mancini   and Federico Ruggeri  and Paolo Torroni 
DISI, University of Bologna, Bologna, Italy
e.mancini@unibo.it

Abstract

Recent advances in NLP suggest that some tasks, such as argument detection and relation classification, are better framed in a multimodal perspective. We propose multimodal argument mining for argumentative fallacy classification in political debates. To this end, we release the first corpus for multimodal fallacy classification. Our experiments show that the integration of the audio modality leads to superior classification performance. Our findings confirm that framing fallacy classification as a multimodal task is essential to capture paralinguistic aspects of fallacious arguments.

1 Introduction

Recent studies in Argument Mining (AM) mainly focus on semantic textual analysis (Lawrence and Reed, 2019). However, a different line of research has shown the importance of including paralinguistic features in argumentative discourse analysis across a wide variety of domains, including advertisements, news coverage, and legal analytics (Kišiček, 2014; Groarke and Kišiček, 2018) and in cognate tasks such as fake news detection (Ivanov et al., 2023). To evaluate these findings, Multimodal Argument Mining (MAM) emerged to gain a more comprehensive understanding of argumentative discourse via integrating multiple modalities. So far, MAM applications include argument detection, argument component classification, and relation classification (Lippi and Torroni, 2016; Mestre et al., 2021; Mancini et al., 2022; Mestre et al., 2023). In contrast, argumentative fallacy classification (Goffredo et al., 2022) has yet to be explored.

While not covering all fallacy types comprehensively, Kišiček (2020) analyzes political discourse to show the connection between human sound, the paralinguistic component of fallacious arguments, and their verbal content. In particular, they link stereotypes on accents to the *ad hominem* fallacy,

as politicians use accents with negative stereotypes to mock or discredit opponents during election campaigns. Likewise, a staccato speech rhythm can be linked to the *appeal to authority* fallacy since it is associated with strictness, authority, and dominance, and prosodic elements emphasizing anger can increase the persuasive impact of *appeal to emotion* fallacies (Kišiček, 2020). These observations lead us to believe that argumentative fallacy classification should be formulated as a multimodal task in the context of political debates.

To tackle this new task, we introduce MM-USED-fallacy, the first corpus for multimodal argumentative fallacy classification. We extend the USED-fallacy¹ corpus (Goffredo et al., 2022) by integrating the audio modality. We follow the methodology described in Mancini et al. (2022) to align audio recordings to annotated debate transcripts. The new corpus contains 1,891 annotated text-audio pairs.

In our evaluation, we extend the multimodal architecture of Mancini et al. (2022) by including state-of-the-art unimodal encoding models, such as pre-trained transformers, that are suitable for low-resource scenarios. Our findings confirm that the combination of text and audio leads to superior classification performance for several models, corroborating our initial hypothesis on argumentative fallacy classification. We release our corpus and code in a public repository.²

2 Related Work

Several studies have targeted political debates in AM (Lippi and Torroni, 2016; Mancini et al., 2022; Mestre et al., 2023), inspired by the recent findings on the importance of paralinguistic components like prosodic features for argument detection (Ben-

¹Henceforth, USED is shorthand for USElecDeb60to16.

²<https://github.com/lt-nlp-lab-unibo/multimodal-am-fallacy>

lamine et al., 2015). Most notably, Mancini et al. (2022) and Mestre et al. (2023) introduced two independent extensions of USED (Haddadan et al., 2019), the US presidential election corpus. These extensions represent the largest to-date multimodal corpora for AM. Similar efforts have been conducted in the context of fake news detection. In particular, Ivanov et al. (2023) observed superior classification performance in several tasks, such as the detection of check-worthy claims, when following a multimodal problem formulation. While the existing studies on political debates have targeted a wide set of AM tasks, the automated analysis of argumentative fallacies has never been explored from a multimodal perspective.

In the context of multimodal deep learning, it is a standard approach to combine unimodal models via fusion techniques (Toto et al., 2021). This is also the case for MAM (Mancini et al., 2022; Mestre et al., 2023). In this work, we extend the methodology introduced by Mancini et al. (2022) to include state-of-the-art models for text encoding, such as RoBERTa (Liu et al., 2019) and SBERT (Reimers and Gurevych, 2019), and audio encoding like CLAP (Wu et al., 2022).

3 Data

3.1 Preliminaries

The term *fallacy* indicates a deceptive, misleading, or generally invalid argument (Hamblin, 1970; Walton, 1987). The USED-fallacy corpus annotates six categories of fallacy. *Appeal to Authority* refers to the use of an expert’s opinion as evidence to back up an argument. An *Ad Hominem* fallacy is characterized by an excessive attack on an arguer’s position. *Appeal to Emotion* usually involves the use of loaded language. *False Cause* regards the misinterpretation of correlation as causation. *Slogans* are brief and striking phrases used to evoke excitement. *Slippery Slope* is an argument that claims exaggerated outcomes for a given action. Table 1 shows examples of annotated fallacies in USED-fallacy.

Since fallacies have various formats and styles, they can span over one or multiple sentences, and may or may not share sentence boundaries. For example, *Appeal to Authority* and *False Cause* typically span over multiple sentences, whereas *Ad Hominem* could be limited to just a few words.

USED (Haddadan et al., 2019) contains annotated transcripts of US presidential debates aired

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

Table 1: Examples of annotated fallacies.

Annotated Element	Description
Dialogue	a debate portion in which the fallacy is present
Snippet	the annotated fallacy in the dialogue
Fallacy	the label of the fallacy
Subcategory	the subcategory of the fallacy
Component Text	the component text in which the fallacy is found or the closest one
Component Label	the type of component
Relation Label	the relation type between the component and the fallacy

Table 2: Annotations in USED-fallacy.

between 1960 and 2016. USED-fallacy (Goffredo et al., 2022) extends USED by annotating fallacies. Table 2 shows the structure of USED-fallacy. It is worth noticing that annotations are at span level and do not always coincide with full sentences.

3.2 Corpus creation

Our corpus, MM-USED-fallacy, integrates MM-USED and USED-fallacy. To build it, we leverage two resources: **span-level** fallacy snippet and argument component annotations from USED-fallacy and **sentence-to-audio** alignment timestamps from MM-USED. Given that the two resources present a mismatch in granularity, we map span-level annotations from USED-fallacy to the sentence level and then align them with audio recordings. Alternatively, one could rely on text/audio alignments at the word level. However, that would require producing a new resource. Since our aim is to evaluate fallacy classification from a multimodal perspective, we decided to build as much as possible on what was available. We leave word-level text-to-audio alignment to future work. Appendix B includes more details about the resources used for

Fallacy	USED-fallacy	MM-USED-fallacy
Appeal to Emotion	1,427	1,102
Appeal to Authority	416	326
Ad Hominem	279	157
False Cause	179	154
Slippery Slope	118	102
Slogans	78	50
Total	2,497	1,891

Table 3: Number of samples in USED-fallacy and MM-USED-fallacy.

mapping between span- and sentence-level annotations and text/audio alignment.

Span-to-sentence mapping. We label a (sequence of) sentence(s) to a fallacy category or argument component type if they overlap with a span labeled as such.

Audio alignment. We use the retrieval tool released by Mancini et al. (2022) to download the audio files corresponding to the debates annotated in USED-fallacy. In this process, we exclude three debates from the corpus for which no audio recording is found (see Appendix C for more details). We then use the text-to-audio alignment timestamps (see Appendix B) to extract audio clips corresponding to the dialogues, the fallacy snippets, and the argumentative components. We perform the alignment via fuzzy string matching since we notice slight transcript mismatches between USED-fallacy and MM-USED due to different text preprocessing. In this process, we also observe a few inconsistencies between USED-fallacy and MM-USED, amounting to duplicate sentences, punctuation differences, and text segments missing for USED-fallacy. For this reason, we perform additional text processing steps and remove misaligned samples (see Appendix D for more details).

Our final corpus, MM-USED-fallacy, comprises 1891 text-audio pairs on 28 debates. Table 3 compares MM-USED-fallacy with USED-fallacy.

3.3 Corpus statistics

Table 4 reports sentence and audio distribution per dialogue, fallacy snippet, and argumentative component in MM-USED-fallacy. We observe that fallacy snippet length varies between one sentence and eight, indicating high annotation variability.

The distribution of snippet length across fallacy categories is shown in Table 5. We observe that 409 (21.6%) fallacy snippets are the span level. We remark that these snippets are mapped to sentence-

	Min	Max	Mean	Std
Text				
Dialogue	16	449	120.92	84.44
Snippet	1	8	1.33	0.82
Component	1	6	1.02	0.23
Audio (in seconds)				
Dialogue	91.28	4271.40	608.68	582.98
Snippet	0.32	74.32	10.40	7.46
Component	0.32	49.68	8.98	6.51

Table 4: Text and audio length distribution (unit: sentences).

Fallacy	Span	Length							
		1	2	3	4	5	6	7	8
Appeal to Emotion	330	482	168	51	27	30	-	5	9
Appeal to Authority	18	162	79	17	30	7	5	8	-
Ad Hominem	15	55	38	13	15	15	6	-	-
False Cause	6	58	26	17	30	-	-	8	9
Slippery Slope	5	52	24	16	-	5	-	-	-
Slogans	35	15	-	-	-	-	-	-	-
Total	409	824	335	114	102	57	11	21	18

Table 5: Snippet sentence length distribution per category (unit: sentences).

level annotations. Among the 658 snippets spanning over two or more sentences, 578 ($\sim 88\%$) are aligned with an exact match. Notably, the fallacy categories of *Appeal To Emotion*, *Appeal to Authority*, and *False Cause* have the longest snippets. *Appeal To Emotion* is also the most frequent class. In contrast, the *Slogan* category has the smallest number of samples. Lastly, some fallacy categories are not present in all debates. We expect that the low representation of certain categories across the debates will have a negative impact on classification accuracy.

Additional insights from the data show that the span length varies between a minimum of 1 word and a maximum of 57 words. Furthermore, 24 sentences, representing 2.2% of the entire dataset, have multiple associated snippets. This small fraction suggests that the agreement statistics closely align with those reported by Goffredo et al. (2022), with minor variations attributed to this subset. Considering the limited occurrence of sentences with multiple associated snippets relative to the total number of snippets in the dataset, we believe that this form of annotation mapping does not introduce significant drawbacks.

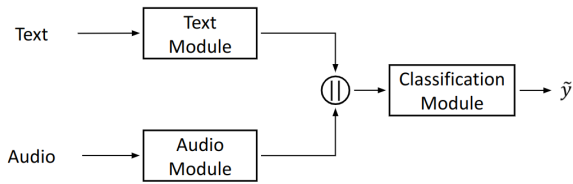


Figure 1: The schema for multimodal argument mining of (Mancini et al., 2022).

4 Experimental Setup

We frame argumentative fallacy classification as a multi-class sentence classification task. We evaluate models via leave-one-out cross-validation, totaling 28 individual model runs. In particular, we build folds such that, at each iteration, each debate is either in the left-out test split or in the remaining splits.

We experiment with the multimodal architecture presented in Mancini et al. (2022) (Figure 1). The text module comprises a pre-trained text embedding model and a dropout layer on top of it. The audio module consists of a pre-trained audio embedding model, a BiLSTM layer, and a dropout layer. The output of the text and audio modules is concatenated and fed to the classification module, defined as a stack of dense layers. We extend (Mancini et al., 2022) by exploring two audio signal encoding and three text encoding models. For audio, we consider Wav2vec (Schneider et al., 2019) and CLAP (Wu et al., 2022), while for text, we use pre-trained BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and SBERT (Reimers and Gurevych, 2019). We also experimented with pre-trained text embeddings without fine-tuning but obtained unsatisfactory results (not reported).

We train all models using cross-entropy loss and Adam (Kingma and Ba, 2015) for optimization. See Appendix E from more details.

To assess the contribution of each modality, we consider three input configurations: *text-only*, *audio-only*, and *text-audio*. We address main fallacy categories only, leaving sub-categories for future work. To limit model overfitting on certain fallacy categories, we remove duplicate snippets, totaling 1063 unique dialogue-snippet pairs.

5 Results

Table 6 reports the macro f1-score for each fallacy category, averaged across all model runs. The text-audio setting leads to performance improve-

ment over text-only (up to 8 points) for BERT (p -value < 0.05) and RoBERTa (p -value > 0.05), independently of the choice of audio model. In contrast, we observe a significant performance drop for SBERT when adding audio.

The observed performance improvements are not equally distributed over fallacy types. For instance, text-only models achieve superior or equal f1-score for the Slogans (S) category. One reason for that could be the low number of examples in this category. Another reason could be that, according to linguistic analyses, slogans primarily rely on figurative language (Dubovičienė and Skorupa, 2014), whereas prosodic features have minimal impact on them (Skorupa and Dubovičienė, 2015).

Classifying fallacies at the sentence level may represent an additional challenge for the audio modality. As shown in Table 5, a notable amount of samples in categories like Appeal to Emotion (AE) and Slogans (S) are annotated at the span level in USED-fallacy. Nonetheless, our findings suggest that audio information is relevant to certain fallacy categories, indicating that this type of problem may benefit from the inclusion of audio features. Determining how to leverage audio-text information remains an open challenge, as the introduction of audio induces variations across different classes, necessitating further studies.

Our investigation into the sensitivity of our results to individual speakers and the influence of audio features on fallacy classification revealed insightful nuances across different debates. While our dataset structure, comprising pairs of candidates per debate, prevented us from pinpointing specific speakers benefiting most from audio features, we observed significant variations across folds and debates. Notably, recent debates tended to show a more pronounced benefit from audio features, possibly due to improved audio quality or the predominance of newer audio files in the training data for audio feature extractors. Specifically, debates such as *1980-Carter/Reagan*, *2004-Bush/Kerry*, and *2008-McCain/Obama* notably benefited from the inclusion of audio features. In other debates, like *1976-Carter/Ford* and *1996-Clinton/Dole*, integrating audio features led to decreased classification accuracy, while debates like *1960-Kennedy/Nixon* yielded mixed results. These findings underscore the complex interplay between speaker characteristics, debate context, and the utility of audio features in enhancing fallacy classification, highlighting the need for further investigation.

	AE	AA	AH	FC	SS	S	Avg ($\bar{x} \pm \sigma$)
Text-only							
BERT	.70	.45	.15	.28	.22	.06	.32 \pm .13
RoBERTa	.53	.50	.32	.29	.30	.17	.38 \pm .18
SBERT	.54	.39	.23	.27	.20	.04	.31 \pm .18
Audio-only							
Wav2Vec	.45	.05	.06	.08	.08	.03	.13 \pm .07
CLAP	.29	.17	.03	.03	.09	.00	.12 \pm .08
Text-Audio							
BERT + Wav2Vec	.80	.50	.13	.35	.23	.04	<u>.40</u> \pm .17
BERT + CLAP	.77	.44	.11	.31	.28	.01	.36 \pm .17
RoBERTa + Wav2Vec	.70	.44	.16	.41	.30	.12	.39 \pm .19
RoBERTa + CLAP	.74	.45	.23	.37	.31	.12	.40 \pm .19
SBERT + Wav2Vec	.45	.29	.27	.21	.11	.04	.23 \pm .11
SBERT + CLAP	.44	.32	.20	.25	.17	.04	.24 \pm .10
Baselines							
Majority	.79	.00	.00	.00	.00	.00	.20 \pm .17
Random	.33	.15	.08	.05	.03	.03	.12 \pm .05

Table 6: Result on MM-USED-fallacy. AE: *Appeal to Emotion*, AA: *Appeal to Authority*, AH: *Ad Hominem*, FC: *False Cause*, SS: *Slippery Slope*, S: *Slogans*. In bold the best model, underlined the second-best model.

6 Conclusion

We posit that argumentative fallacy classification should be framed as a multimodal task. To empirically evaluate our hypothesis, we build the first dataset for multimodal argumentative fallacy detection, MM-USED-fallacy. Our results show that the integration of audio modality is indeed beneficial, observing significant performance improvement (4-8 f1-score percentage points) in a variety of model architectures.

Our findings are coherent with recent studies in argument mining (Mestre et al., 2023; Mancini et al., 2022) and fake news detection (Ivanov et al., 2023). We believe that a multimodal formulation should affect the experimental setting starting from data collection, if possible, so as to capture several audio properties like non-verbal features (Kišiček, 2020) in addition to prosodic ones.

We believe that the multimodal resource that we provide has significant potential for enabling further experimentation. Some possibilities are addressing fallacy subcategories, experimenting with other tasks like argumentative fallacy detection, evaluating the importance of argumentative components (Goffredo et al., 2022), and employing novel multimodal architectures. Moreover, to gain a deeper understanding of our results, employing interpretability techniques designed to emphasize the significance of paralinguistic elements over linguistic elements in prediction may be beneficial.

Limitations

Datasets. This study is based on a single dataset. Moreover, not all the text in USED-fallacy (Goffredo et al., 2022) could be used, since in some cases audio-to-text alignment was unsuccessful (see Section 3). For this reason, some of the fallacies annotated in USED-fallacy are lost in MM-USED-fallacy.

Annotations. We argue that fallacy classification should be framed as a multimodal task. However, the annotations utilized in this research were derived from those defined in USED-fallacy (Goffredo et al., 2022) based on the textual content only. Such annotations are likely to disregard the potential insights coming from the acoustic elements of the debates. In order to take into account all sound-related cues, a new annotation of fallacies should be carried out from scratch, using a new set of guidelines.

Experimentation. Like prior art (Goffredo et al., 2022), this study is also limited to argumentative fallacy classification, and to a few selected models for text and audio embedding. For more robust results, the study could include other tasks, like fallacy detection, and other text/audio embedding modes, as well as different alignment architectures like that introduced by (Ivanov et al., 2023).

Ethics Statement

The automatic detection of argumentative fallacies could help gain insights into the persuasive techniques employed by politicians. This could have a positive impact on society by promoting critical thinking and informed decision-making among the public or as a support to educational initiatives at school, and ultimately a more robust democratic process.

We believe that this work in itself is not harmful to anyone. Our primary focus is on improving the understanding and detection of argumentative fallacies, not on promoting negativity or harm toward individuals or groups. We do not take any stance on the content of the debates or on the individuals involved or mentioned in them. All data we used was publicly available.

While we acknowledge that the dataset we rely upon may have inherent biases, we have taken measures to mitigate them to the best of our abilities. However, we understand that biases can exist in any dataset, and we are committed to transparency

and accountability. By making our work public, we invite scrutiny and analysis from the research community, enabling future work to identify and correct any biases or errors that may be present. This iterative process helps to refine and improve the accuracy and fairness of our methodology over time.

References

- M. Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. [Emotions in argumentation: an empirical evaluation](#). In *IJCAI*, pages 156–163. AAAI Press.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.
- Tatjana Dubovičienė and Pavel Skorupa. 2014. [The analysis of some stylistic features of english advertising slogans](#). *Žmogus ir Žodis*, 16:61–75.
- Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. [Fallacious argument classification in political debates](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4143–4149. ijcai.org.
- 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. 1970. *Fallacies*. Vale Press, Newport News, Va.
- Petar Ivanov, Ivan Koychev, Momchil Hardalov, and Preslav Nakov. 2023. [Detecting check-worthy claims in political debates, speeches, and interviews using audio data](#). *CoRR*, abs/2306.05535.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *ICLR*.
- 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. 2020. [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.
- John Lawrence and Chris Reed. 2019. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- Marco Lippi and Paolo Torrioni. 2016. [Argument mining from speech: Detecting claims in political debates](#). In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI’16*, 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](#). *CoRR*, abs/1907.11692.
- Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torrioni. 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.
- Rafael Mestre, Stuart 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. 2021. [M-arg: Multimodal argument mining dataset for political debates with audio and transcripts](#). In *ArgMining@EMNLP*, pages 78–88. Association for Computational Linguistics.
- 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.
- Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. [wav2vec: Unsupervised pre-training for speech recognition](#). In *INTERSPEECH*, pages 3465–3469. ISCA.
- Pavel Skorupa and Tatjana Dubovičienė. 2015. [Linguistic characteristics of commercial and social advertising slogans](#). *Coactivity: Philology, Educology*, 23:108–118.

- Ermal Toto, ML Tlachac, and Elke A. Rundensteiner. 2021. [Audibert: A deep transfer learning multimodal classification framework for depression screening](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 4145–4154, New York, NY, USA. Association for Computing Machinery.
- Douglas N. Walton. 1987. *Informal Fallacies: Towards a Theory of Argument Criticisms*. John Benjamins, Philadelphia.
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. [Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation](#). *CoRR*, abs/2211.06687.

A Terminology

In this section, we provide formal definitions for the key terminology utilized throughout the paper to facilitate clarity and understanding.

Snippet. The term *snippet* refers to the annotated fallacy within the dialogue.

Component. The term *component* denotes an argumentative element such as *claim* or *premise*.

Component Text. It signifies the text containing the component or its nearest representation within the discourse structure.

Component Label. This term indicates the type assigned to a particular argumentative component within the discourse framework.

Span. In the context of this study, *span* and *span-level annotations* refer to the characterization of fallacies by groups of words. These groups may not necessarily form complete sentences but could extend over multiple sentences or constitute substrings within a sentence.

Sentence. A *sentence* is defined as a unit of text delimited by a full stop.

B External Resources for Dataset Construction

The resources provided in Mancini et al. (2022) include:

- A file that establishes the correspondence between the debate IDs and the debate recordings available on the PBS NewsHour YouTube channel³. This file also contains timestamps necessary for removing audio content not present in the paired transcripts, such as opening and closing remarks by the moderators.
- Pre-processed transcripts that have undergone several transformations to achieve alignment between the audio files and the text. These transformations include (1) the removal of sentences in the transcripts that do not match the audio file cuts; (2) the removal of metadata (e.g., speaker information); (3) transcripts sentence splitting; (4) the extraction of transcripts corresponding to each 20-minute portion of the audio files.
- JSON files containing the alignment timestamps for each 20-minute audio chunk in each

³<https://www.youtube.com/channel/UC6ZFN9Tx6xh-skXCuRHCDpQ>

debate. These files include the start and end timestamps of each utterance. The alignment was performed at the sentence level, such that each utterance corresponds to one sentence in the debate.

C Dataset Pre-Processing Details

According to Goffredo et al. (2022), only 31 out of the 39 debates in the USED corpus are annotated with fallacies. Additionally, USED-fallacy contains a new debate w.r.t. USED and MM-USED, namely the third 2016 presidential debate between Clinton and Trump. Moreover, we exclude three debates from USED-fallacy, due to discrepancies between the audio recording and the corresponding transcripts. These debates are the first 1998 parliamentary debate between Bush and Dukakis and the first two 2016 presidential debates between Clinton and Trump. For the same reason, we also exclude the second section of the first 1992 debate between Clinton, Bush, and Perot.

To simplify audio-to-text alignment, we merge the JSON alignment files provided in Mancini et al. (2022) to obtain a single alignment file for each debate. We adjust the start and end timestamps of the files after the first chunk. Specifically, we add a duration equal to 20 minutes multiplied by the identifier of the chunk. For example, the timestamps of the second alignment file (`chunk_id = 1`) of a debate are shifted by 20 minutes.

D Dataset Cleaning and Additional Alignment Operations

We notice that 10 snippets and 5 components are missing from their corresponding dialogues. Thus, we remove the corresponding samples from the corpus. Furthermore, we discover that when the first sentence of the dialogue is a duplicate sentence in the alignment file (e.g., *Ok* or *Thank you*), the timestamp that is associated with the sentence always corresponds to the last occurrence of that sentence in the alignment. We correct such misalignments manually.

E Training Details

The primary focus of our work lies not in achieving absolute performance but rather in facilitating a comparative analysis across diverse modalities. In line with this objective, we determine hyperparameters grounded in our prior experience as outlined in Table 7. For all models, we employ class weights

to manage training data imbalance. Each model is trained on a single GPU (NVIDIA 2080Ti) with 12 GB dedicated memory in less than 24 hours.

Modality	Text Model	Audio Model	BS	Epochs	Seed	LR
AO	BERT/SBERT	CLAP/Wav2Vec	8	500	15371	5e-05
TA	BERT/SBERT	CLAP/Wav2Vec	8	500	15371	5e-05
TO	RoBERTa	CLAP/Wav2Vec	32	100	15371	5e-05

Table 7: Hyper-parameters Configurations. *Modality*: Input Modality where *AO*, *TA*, and *TO* refer to *audio-only*, *text-audio* and *text-only* respectively, *BS*: Batch Size, *Epochs*: Number of Training Epochs, *Seed*: Random Seed, *LR*: Learning Rate.