Multimodal Argument Mining: A Case Study in Political Debates

Eleonora Mancini and **Federico Ruggeri**✉ and **Andrea Galassi** and **Paolo Torroni**
DISI, University of Bologna, Bologna, Italy
federico.ruggeri6@unibo.it

Abstract

We propose a study on multimodal argument mining in the domain of political debates. We collate and extend existing corpora and provide an initial empirical study on multimodal architectures, with a special emphasis on input encoding methods. Our results provide interesting indications about future directions in this important domain.

1 Introduction

Argument mining (AM) aims to extract arguments and their relations from natural language sources (Lippi and Torroni, 2016b). Performing AM usually entails tackling one or more tasks like argumentative component detection and classification, link prediction, relation classification, or stance classification (Lawrence and Reed, 2020) in a particular domain of interest. Among the many areas and genres where AM was investigated, the political domain allows for intuitive applications with the final aim of detecting fallacies, persuasiveness degree (Cano-Basave and He, 2016), truthfulness (Nakov et al., 2018; Kopev et al., 2019) and coherence in the candidate's argumentation (Cabrio and Villata, 2018; Lippi and Torroni, 2016a), or summarizing the candidate's positions (Vilares and He, 2017). So far, most of AM research has focused on textual inputs. Political debates and speeches have been no exception. However, differently from other domains, this particular one is especially rich in audio input sources. This could be important, since the audio input, in addition to text, may leverage the exploitation of para-linguistic cues related to the argumentation process, improving the performance of argumentative component detection and other AM tasks (Lippi and Torroni, 2016a; Villata et al., 2017; Polo et al., 2016). To date, partly owing to the scarcity of non-textual corpora for AM (Haddadan et al., 2019), only a couple of attempts have been made in this direction. Conversely, outside of AM, in the broader area of Natural Language Processing (NLP), Multimodal Deep Learning (MMDL) is attracting growing interest, also owing to remarkable progress made in the field. Current research in MMDL focuses on advanced input representations and fusion solutions. These include end-to-end architectures fully based on transfer learning for input representation (Toto et al., 2021) and attention-based architectures for efficient input management (Lian et al., 2019; Tsai et al., 2019; Gu et al., 2018). These noteworthy developments suggest that time is ripe to rethink multimodal AM in light of the latest findings in multimodal NLP research.

In spite of a wide availability of raw audio sources, processing and annotating good quality data can be very costly. To the best of our knowledge, the only two multimodal AM corpora on political speeches are UKDebates (Lippi and Torroni, 2016a), which addresses the task of claim detection, and M-Arg (Mestre et al., 2021), which focuses on argumentative relations between sentences. These are small-sized corpora where a handful of speakers debate in one or a few occasions over a year's time span. On the other hand, USElecDeb60To16 is a corpus curated by Haddadan et al. (2019), where a significant number of US presidential candidates debate over a time span of several decades. However, it only contains annotated transcripts, with no link to the audio source.

In an effort to push the envelope in multimodal AM, with this work we release MM-USElecDeb60To16, an extended version of the USElecDeb60To16 corpus, where the text input is complemented by and aligned to the audio input. At the time of writing, this is the largest multi-modal resource for AM in the domain of political debates, as well as the one with the largest number of speakers, and longest time span covered. These features make MM-USElecDeb60To16 a particularly challenging corpus, since para-linguistic cues are very much speaker-dependent (Lippi and Torroni, 2016a) and political communication, argumentation, and language have greatly evolved in such a long time span (Haddadan et al., 2019).

Alongside this new resource, we offer a preliminary but rigorous and reproducible experimental study of multimodal AM in political debates. Our benchmarks are all the relevant corpora available: UKDebates, M-Arg, and MM-USElecDeb60To16. We build and compare architectures inspired to proposals from literature, in order to study the effect of changing the encoding of the audio input. In particular, we compare the more traditional feature-based audio encoding, with a more advanced input encoding technique that builds on recent findings in MMDL for NLP. Our results indicate that the encoding of the audio input has a noticeable effect on performance, but they also suggest that a better fusion of textual and audio input encodings and more advanced architectural solutions might be needed in order to make progress in the more challenging tasks and corpora.

The paper is structured as follows. In Section 2 we overview related work in multimodal AM and multimodal deep learning (MMDL), with a focus on architectures for text and audio processing. In Section 3 we discuss corpora and in Section 4 we define the AM tasks addressed. Section 5 presents the experimental setup and describes models, input encodings and training. Section 6 discusses the results of our experimental study. Section 7 concludes. In appendix we report all the information needed for reproducibility. The corpus and the code are publicly released.[1]

2 Related Work

There exists a strong connection between the argumentation process and the emotions felt by people involved in such a process (Benlamine et al., 2015). This observation motivated the hypothesis that para-linguistic elements encoded in the audio data are significant indicators that might aid identify arguments made in a debate. Recent studies confirmed this hypothesis. In particular, in the domain of political debates, Lippi and Torroni (2016a) presented a case study in AM from speech using a televised debate from the 2015 UK political elections. They built a first-of-a-kind political debate corpus by annotating arguments uttered by three prime ministerial candidates, and showed that audio features helped claim detection when used as input to a Support Vector Machine (SVM) classifier together with their textual transcript. More recently, Mestre et al. (2021) built the M-Arg corpus, which consists in 4,104 labelled pairs of sentences selected from debates of the 2020 US political elections. They experimented on this new corpus using a different multimodal input model. Outside of political speeches, a corpus that couples transcript and audio of several debates was developed by Mirkin et al. (2018a,b). However, differently from the previous corpora, here non-political debates are carried out by paid actors on a set of controversial topics taken from the iDebate web site.

To the best of our knowledge, at least in political debates, multimodal AM has not been further explored. Reasons for that lie partly in the difficulty and heterogeneity of AM tasks, partly in the scarcity of multimodal data for AM, partly in the inherent challenges of multimodal deep learning (MMDL). One such challenge is in endowing models with the ability to digest and actually benefit from different, complementary modalities. In this respect, the works by Lippi and Torroni (2016a), Villata et al. (2017) and Mestre et al. (2021) could be viewed as proofs-of-concept of the potential of multimodality in AM. They used mostly traditional methods for categorising data and encoding audio, such as SVM classifiers and MFCCs. Like most other studies in AM, they were carried out on a single corpus and a specific task.

Recent MMDL solutions suggest a number of promising directions for improvement. These include full transfer learning-based frameworks to alleviate the problem of multimodal data shortage (Zhang et al., 2022; Naderi et al., 2019; Harati et al., 2018) and attention mechanisms to handle interactions among and between different modalities (Lian et al., 2019; Tsai et al., 2019; Gu et al., 2018). For example, AudiBERT (Toto et al., 2021), a recent MMDL architecture, integrates pre-trained text and audio models via a dual self-attention mechanism. In our work, we examine the architectural designs presented in earlier studies (Lippi and Torroni, 2016a; Mestre et al., 2021) and also suggest a multimodal architecture comparable to AudiBERT, based on text and audio embedding taken from pre-trained models like GloVe (Pennington et al., 2014), BERT (Devlin et al., 2019)

[1] https://github.com/federicoruggeri/multimodal-am/

and Wav2Vec (Schneider et al., 2019).

3 Data

We experiment on three different debate corpora, designed to address four separate but strictly correlated argument mining tasks. Table 1 summarizes the corpora's key figures.

3.1 UKDebates Corpus

UKDebates, by Lippi and Torroni (2016a), was the first corpus released for multimodal argument mining. Its context is the UK Prime Ministerial elections of 2015. It is based on the two-hour debate aired by Sky News on April 2, 2015 and it comprises the audio sequences of 3 candidates: David Cameron, Nick Clegg, and Ed Miliband. UKDebates contains 386 audio samples (122 for David Cameron, 104 for Nick Clegg, 160 for Ed Miliband) of varying length, accompanied by a human-built transcript. Two domain experts annotated the collected transcripts for the task of claim detection, by labeling each sentence as containing or not containing a claim. Regarding Inter-Annotator Agreement (IAA), the authors report $\kappa = 0.53$, "fair to good" agreement. Because audio features are markedly speaker-dependent, Lippi and Torroni (2016a) addresses the claim detection (CD) task for each individual politician candidate in turn. The authors report a F1-score in the range of $\sim$59-62%.

3.2 M-Arg Corpus

M-Arg, by Mestre et al. (2021), is built around the 2020 US Presidential debates. The debates involve 5 different speakers (4 candidates and a moderator) and are related to 18 topics. A carefully designed crowd-sourcing exercise resulted in 4,104 labelled sentence pairs for the task of argumentative relation detection. In particular, each sentence pair was labeled as *support*, *attack*, or *neither*. To account for crowd-workers' annotation quality, each label is enriched with an *annotator agreement confidence* γ. A smaller but higher-quality subset of M-Arg is thus obtained by only selecting the links with confidence $\gamma \geq 0.85$. The price of this reduction in the annotations' noise is a reduced size of the dataset, which results in harder training and an IAA of $\alpha = 0.43$. Mestre et al. (2021) report a macro F1-score of 22.5% and 11.0% for the argumentative relation classification (ARC) task regarding the full corpus and the ($\gamma \geq 0.85$) subset, respectively. The macro F1-score regards the *support* and *attack* labels only.

3.3 MM-USElecDeb60to16 Corpus

USElecDeb60To16, by Haddadan et al. (2019), is the largest collection of annotated textual documents for argument mining in the political debates domain. It contains presidential candidates' debate transcripts aired from 1960 to 2016. Annotations are at the sentence level. Each sentence is labeled as a claim, a premise, or neither of them. The authors used this corpus to address the argumentative sentence detection (ASD) and argumentative component classification (ACC) tasks. Regarding IAA, they report a $\kappa = 0.57$ (moderate agreement) for ASD and of $\kappa = 0.40$ (fair agreement) for ACC. As for classification performance, Haddadan et al. report a macro F1-score of 73.0% and 76.95% for the ASD and ACC, respectively.

We build MM-USElecDeb60To16 by augmenting USElecDeb60To16 with the audio modality. We remark that we do not add any additional label, nor we modify existing ones. We obtained the debates audio files from the PBS NewsHour YouTube channel.[2] Before aligning transcripts with corresponding audio files, we carried out a preliminary pre-processing phase. First, we manually trimmed audio files to remove content that is not included in the paired transcripts, such as some of the opening and closing remark of the moderators. In some cases, audio files can contain cuts spanning from a few seconds to several minutes. We removed the transcripts' sentences that were matched to these cuts. Second, we removed transcripts that did not match their paired audio files or incomplete ones. Third, we removed metadata like the speaker's information from each transcript to avoid spurious alignments. Fourth, we tokenized transcripts; thus, the resulting transcripts contain one sentence per line. See Appendix A for additional details.

After pre-processing, we split each audio file into 20-minute chunks to improve the alignment quality. We manually extract the transcripts' text corresponding to the created audio files. We used Aeneas[3] to automatically retrieve the start and end timestamps of each utterance. Lastly, we post-processed our corpus by removing (i) sentences misaligned with their audio sample (ii) sentences not matching any of the aligned utterances

[2] https://www.youtube.com/channel/UC6ZFN9Tx6xh-skXCuRHCDpQ

[3] https://github.com/readbeyond/aeneas/

Corpus	Sentences	Debates	Speakers	Years	Class Distribution	Task(s)
UKDebate (Lippi and Torroni, 2016a)	386	1	3	2015	152 claim, 234 not-claim	CD
M-Arg (Mestre et al., 2021)	4,104 pairs	5	4	2020	120 attack, 384 support, 3600 neither	ARC
M-Arg ($\gamma \geq 0.85$) (Mestre et al., 2021)	2,443 pairs	5	4	2020	29 attack, 132 support, 2282 neither	ARC
MM-USElecDeb60to16 (Ours)	26,781	39	26	1960-2016	10,882 claim, 9,683 premise, 6,226 not-arg	ASD, ACC

Table 1: Corpora for multimodal argument mining. For M-Arg, we also consider the corpus version where samples have high annotation confidence γ. The acronyms used in column Task are spelled out in Section 4.

(e.g., transcription tags like "applause") (iii) non-argumentative duplicated sentences, such as *Thank You* or *Ok*. Finally, we verified the quality of the alignments by spot checks. In particular, we checked several different parts of each debate, and no major misalignments were found.

As a result of the mentioned pre- and post-processing phases, we removed about 2,000 samples from the original USElecDeb60to16 corpus. The resulting MM-USElecDeb60to16 corpus contains 26,791 annotated textual sentences and their corresponding audio samples.

Our corpus differs from previous multimodal AM corpora in terms of size, variety and annotation quality. First off, it is the largest multimodal AM corpus to date, by a significant margin. Second, it offers a wider range of speakers over a much longer time span (1976-2016), possibly paving the way to new research perspectives, such as the analysis of the evolution of political communication, argumentation and language over time. The greater number of different speakers could also facilitate the creation of more robust classification models. Finally, the corpus includes expert annotations, as opposed to crowd-sourced ones.

4 Methodology

We consider four distinct classification tasks:

- **Argumentative Sentence Detection (ASD)**: an input sentence x is classified as containing an argument (*arg*), or not containing an argument (*not-arg*);
- **Argumentative Component Classification (ACC)**: an argumentative sentence x is classified as containing a *claim* or a *premise*;
- **Claim Detection (CD)**: a sentence x is classified as containing a *claim* or not containing a claim (*not-claim*);
- **Argumentative Relation Classification (ARC)**: a pair of sentences x_i and x_j is classified as yielding an argumentative relation $x_i \rightarrow x_j$ of *support*, *attack*, or *neither* (if no argumentative relation exists).

Each input is characterized by two modalities: the textual input x_t and the audio input x_a. To assess the impact of each modality, we consider three distinct input configurations: *text-only* (TO), *audio-only* (AO), and *text-audio* (TA), where both modalities are given as input.

5 Experimental Setup

We define a reproducible and robust experimental setup to evaluate the contribution of each modality to AM tasks, and to assess the impact of different input representations and classifiers. The limited amount of data, especially in UKDebates and M-Arg, caused our setup to differ in several ways from previous studies. Hence our results are not directly comparable with those published in the relevant literature. Nonetheless, our setting includes all the classifiers used in such studies, in addition to more recent representation techniques.

Regarding UKDebates, Lippi and Torroni (2016a) address the CD task by experimenting on each politician individually. In particular, the authors evaluate a multimodal SVM classifier via a

10-fold cross-validation routine. In contrast, we evaluate our models on all speaker sentences via a repeated 5-fold cross-validation routine. Such a design choice was made to curb the high variance usually observed in a model's performance when neural models are trained with little data (Bengio, 2012). We set the number of repetitions to 3.

For the same reason, we evaluate our models on M-Arg via a repeated 5-fold cross-validation routine. We set the number of repetitions to 3. Our approach differs from the one proposed by Mestre et al. (2021), where the corpus is divided into train and validation splits.

For MM-USElecDeb60to16 we follow the same experimental setup as in (Haddadan et al., 2019). Despite the different number of samples, we keep the same train, validation, and test splits proposed for the original corpus. We define a repeated training and evaluation routine for model benchmark, setting the number of repetitions to 3. See Appendix B for additional details on our experimental setting, number of samples and data splitting.

5.1 Models

We defined three models, according to the high-level schema illustrated in Figure 1. In all our models, each modality is processed separately by either a *text module* or an *audio module*. Each module is part of a classification model defined for a particular input modality. Different input configurations use different modules. The TO and AO input configurations only consider the text module or the audio module, respectively. In the TA multimodal setting, instead, the outputs of the two modules are concatenated and passed through a final *classification module*. The classification module receives the encoded representation of one or multiple modalities according to the considered input configuration and produces a classification label.

For each model we experiment with two different audio signal encoding methods: a set of widely-adopted spectral features (Rejaibi et al., 2022) and the Wav2vec embeddings (Schneider et al., 2019). Such encoding methods represent a preliminary pre-processing step of the audio signal, which is then passed in input to the audio module.

The models are defined as follows.

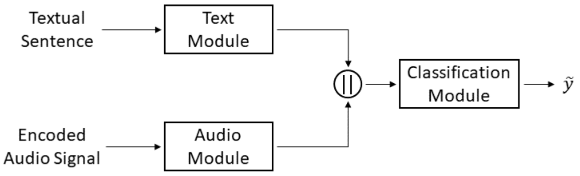

Figure 1: The proposed schema for multimodal argument mining.

- **SVM** follows Lippi and Torroni (2016a). The text module encodes input textual sentences as TF-IDF vectors. The audio module is an identity function, that is, the encoded audio signal remains unaltered. The classification module is an SVM classifier.

- **M-ArgNet** reflects the neural architecture presented in Mestre et al. (2021). The text module is defined by a pre-trained BERT (Devlin et al., 2019) model. The audio module is a stack of CNN layers with a BiLSTM layer on top. The classification module is a MLP.

- **BiLSTM** is a third architecture where the text module comprises a pre-trained GloVe (Pennington et al., 2014) embedding layer to encode input textual sentences and a stack of BiLSTM layers. The audio module is defined by another stack of BiLSTM layers. The classification module is a MLP.

M-ArgNet and BiLSTM, when used with Wav2vec embeddings, resemble AudiBERT (Toto et al., 2021) since they are all based on text and audio embedding taken from pre-trained models.

In addition to the above models, we also consider a weighted random baseline classifier, i.e. **Random**, which acts as a lower bound for each task of interest.

5.2 Audio Representation

In this section, we provide additional details regarding the described audio signal encoding methods. The first method, denoted as *feature-based* encoding, is a set of widely-adopted spectral features (Rejaibi et al., 2022), such as the Mel-frequency cepstral coefficients (MFCCs), spectral centroids, spectral bandwidth, spectral roll-off, spectral contrast and a 12-bit chroma vector. The result of this extraction process is a two-dimensional feature matrix of shape (no. frame, no. features). Following Mestre et al. (2021), we consider 25 MFCCs and 20 other spectral features, for a total of 45 features.[4] Regarding the number of frames, their amount is proportional to the duration of the audio

[4] We used the `librosa` library with default parameters.

signal. In our experimental setup it is in the order of hundreds. To reduce the number of frames we adopt average pooling. This applies a moving average with a parametric window size on the frame dimension. We experiment with different window sizes to reduce the computational demand and the number of parameters of our models, without degrading the informative content of the audio signal.

The second method, denoted as *embedding-based* encoding, uses the end-to-end audio encoding neural architecture Wav2vec (Schneider et al., 2019).[5] In particular, we directly extract the pooled embedding vector given by the model. We denote this setting as *embedding-based* encoding. The final size of the representation is a 768-dimensional embedding vector according to the chosen Wav2vec model.

5.3 Optimization

We train our neural models using cross-entropy as the optimization objective and Adam (Kingma and Ba, 2015) as an optimizer. Additionally, we regularize neural models by applying early stopping on the validation loss with patience set to 10 epochs and using dropout (Srivastava et al., 2014).

All models undergo a preliminary hyper-parameters calibration phase. In particular, for each input configuration (i.e., TO, AO, and TA) we calibrate the models to assess the contribution of individual modalities. Additional details about model calibration are reported in Appendix C.

6 Results

We report the classification results on each dataset. Additionally, we perform an ablation study regarding the models trained in the TA configuration to evaluate the contribution of each input modality.

UKDebates Table 2 reports classification results for the CD task on the UKDebates corpus. In particular, we compute the average binary F1-score on the test set for each input configuration and audio encoding method. We provide the F1-score as a customary performance indicator in unbalanced classification situations. We observe that the best-performing input configuration for each model is the TA with embedding-based audio encoding. However, the gap with respect to the TO input configuration is marginal, suggesting that the audio modality is not efficiently handled by the employed models or is not sufficiently informative. Regarding audio encoding, we observe that the embedding-based method leads to better performance than the feature-based approach. This is evident for the SVM classifier, where the TA setting with embedding-based audio encoding leads to an improvement of more than 17 F1-score percentage points compared to its feature-based counterpart.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	66.24	48.62	49.13	46.20	**66.71**
M-ArgNet	67.20	47.20	65.94	50.12	**68.68**
BiLSTM	66.81	45.40	65.29	50.20	**66.84**
Random			40.90		

Table 2: Average binary F1-score regarding the *claim* class on the test set of UKDebates. For each row, we report the best results in bold, second best results are underlined instead.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	14.70	11.96	12.33	14.09	**16.73**
M-ArgNet	16.24	18.45	18.27	8.88	**19.02**
BiLSTM	16.78	9.18	15.89	9.84	**20.21**
Random			2.79		

Table 3: Average macro F1-score concerning the *attack* and *support* classes on the test set of M-Arg ($\gamma \geq 0.85$). For each row, we report the best results in bold, second best results are underlined instead.

M-Arg Table 3 reports the average macro F1-score regarding the *attack* and *support* classes on the test of the M-Arg corpus for the ARC task. We focus on the M-Arg corpus version with annotation confidence $\gamma \geq 0.85$ to consider high-quality examples only. We observe that the TA input configuration with the embedding-based audio representation is the best performing one for all the considered classifiers. In particular, such a configuration outperforms the TO input configuration by 2.03, 2.78 and 3.43 F1-score percentage points for SVM, M-ArgNet, and BiLSTM classifiers, respectively. In contrast, the TA input configuration with feature-based audio representation yields mixed results. More precisely, only the M-ArgNet model outperforms its TO counterpart. This is in agreement with the results obtained in CD on UKDebates. The feature-based AO input configuration is remarkably on par with its TA counterpart.

[5] We use the *facebook/wav2vec2-base-960h* model version.

MM-USElecDeb60to16 Table 4 reports classification performance concerning the ASD and ACC tasks evaluated on the test set of the MM-USElecDeb60to16 corpus. In general, the embedding-based audio encoding appears to perform better than the feature-based one. This agrees with the behaviour observed in the previous experiments. However, we observe that the TA configuration does not always perform better than TO. We hypothesize that the characteristics of this corpus, with multiple speakers spanning several decades, bring in additional challenges that these architectures are not addressing effectively. Concerning ASD, we observe that the TA input configuration is the best performing one for the BiLSTM and the M-ArgNet models. In contrast, the TO input configuration leads to superior performance for the SVM model. Overall, there is no significant performance gap between the TA and TO input configurations. However, the AO input configurations with both audio signal representations are not far behind their TO and TA counterparts. All this suggests that the encoded audio signal is informative to address the task, but the fusion of both modalities is non-trivial depending on the given audio representation. We observe similar behaviours concerning the ACC task. In particular, the TA input configurations do not lead to consistent performance benefits for the employed models. Nonetheless, the AO input configuration with embedding-based audio representation significantly outperforms its feature-based counterpart. These observations confirm a known fact, that merging multiple input modalities is still a major challenge in current multimodal models.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
ASD					
SVM	**67.18**	49.37	49.02	61.20	65.38
M-ArgNet	65.64	52.71	60.89	61.04	**68.38**
BiLSTM	67.19	58.89	**68.57**	60.40	68.23
Random			50.54		
ACC					
SVM	**65.85**	50.19	51.66	58.44	64.75
M-ArgNet	**67.40**	50.05	60.09	65.33	67.38
BiLSTM	65.99	49.58	**66.25**	58.86	65.80
Random			50.51		

Table 4: Average macro F1-score on the test set of MM-USElecDeb60to16. For each row, we report the best results in bold, second best results are underlined.

Discussion The results presented so far warrant the following considerations:

1. Embedding-based audio encoding generally yields better results than feature-based encoding. This is consistent with recent findings in MMDL (Schneider et al., 2019) and confirms that investigating the ramifications of those findings for multimodal AM is a worthwhile endeavour, which should be pursued.

2. The TA input configuration is superior to its TO counterpart, or at least on part with it, in all described corpora. This reinforces our belief that audio can benefit AM tasks. This is also supported by the observed performance of models trained in the AO input configuration. For instance, the performance gap between TO and TA configurations is only ∼2-8 F1-score points for the ASD and ACC tasks in the MM-USElecDeb60to16 corpus.

3. The definition of effective methods for input encoding and fusion represent major challenges of multimodal AM, as observed in our extended case study.

6.1 Ablation Study

To assess the contribution of each individual input modality, we carried out an ablation study on models trained with the TA input configuration, by alternatively masking either input modalities. To this end, we zeroed out the output embedding vector of the input module corresponding to the modality to be masked.

Table 5 reports the results of the ablation study regarding the UKDebates corpus. We observe that the BiLSTM model with feature-based audio representation reaches the same performance in both the TA and TO (i.e., w/o Audio) configurations. This result suggests that the audio modality does not provide informative content in addition to text for the task. From a reversed perspective, the SVM classifier with feature-based audio representation focuses solely on the audio modality. We interpret this as an effect of the difficulty of combining text and audio modalities for the SVM classifier. We observe similar behaviours when considering the embedding-based audio representation as well. In contrast, the M-ArgNet model behaves in line with our initial expectations regarding the ablation study. In particular, the model achieves superior performance compared to the random baseline when one

Input	BiLSTM	SVM	M-ArgNet
Feature-based			
TA	65.29	49.13	65.94
w/o Text	21.00	49.13	46.04
w/o Audio	65.29	0.00	57.02
Embedding-based			
TA	66.84	66.71	68.68
w/o Text	3.81	16.40	11.27
w/o Audio	66.78	0.00	68.48

Table 5: Ablation test regarding TA model configuration on the UKDebates test set.

of the input modalities is removed, while being inferior to the default TA case. The only exception concerns the embedding-based audio representation setting. In this setting, the text modality significantly contributes to the task compared to the audio modality.

Likewise, with the M-Arg corpus (see Table 6), we observe odd results similar to those observed with UKDebates. In particular, the BiLSTM and SVM models show symmetrical effects concerning performance metrics when one of the input modalities is removed. Independently of the audio representation method, the BiLSTM model heavily relies on text information to perform the task. In contrast, the SVM model fails to address the task when audio is removed. This evidence suggest that the way input is encoded also plays an important role in a multimodal model concerning the impact of each modality.

Furthermore, we observe similar issues in the MM-USElecDeb60to16 corpus when addressing the ASD and ACC tasks. Table 7 reports the results of the ablation study concerning the ASD task. Again, we observe that the BiLSTM and SVM models have symmetric behaviours. Additionally, the BiLSTM reaches superior classification performance when removing the audio modality in both audio representation settings. Despite a small improvement, this surprising result suggests that the audio modality might be noisy and, thus, detrimental to the task. This observation is further supported by the low performance achieved when removing the text modality. We observe this phenomenon also in the ACC task as reported in Table 8. In particular, the M-ArgNet with embedding-based audio representation has superior performance when removing the audio modality compared to the default TA input configuration.

Input	BiLSTM	SVM	M-ArgNet
Feature-based			
TA	15.89	12.33	18.27
w/o Text	0.0	12.33	9.21
w/o Audio	10.00	0.00	3.78
Embedding-based			
TA	20.21	16.73	19.02
w/o Text	0.0	9.30	1.16
w/o Audio	12.80	0.00	18.24

Table 6: Ablation test regarding TA model configuration on the M-ARG ($\gamma \geq 0.85$) test set.

Input	BiLSTM	SVM	M-ArgNet
Feature-based			
TA	68.57	49.02	60.89
w/o Text	17.26	49.02	23.11
w/o Audio	69.40	17.26	48.04
Embedding-based			
TA	68.23	65.38	68.38
w/o Text	17.26	61.11	44.35
w/o Audio	68.44	17.26	33.95

Table 7: Ablation test regarding TA model configuration on the MM-USElecDeb60to16 test set for the ASD task.

Input	BiLSTM	SVM	M-ArgNet
Feature-based			
TA	66.25	51.66	60.09
w/o Text	32.71	51.66	44.25
w/o Audio	66.24	32.71	55.25
Embedding-based			
TA	65.80	64.75	67.38
w/o Text	32.71	48.86	33.95
w/o Audio	65.80	33.57	67.57

Table 8: Ablation test regarding TA model configuration on the MM-USElecDeb60to16 test set for the ACC task.

7 Conclusion

Political debates and speeches are an important domain where audio data is abundant. The automated argumentative analysis of such data could leverage a variety of innovative applications and open promising research avenues. Yet, AM research so far has mostly focused on textual transcripts. Motivated by recent advances in MMDL and in an effort to push the envelope in multimodal AM research, we release the largest-to-date multimodal AM dataset. We thus run an empirical study on three multimodal AM datasets differing from one another in many respects like size, topics, annotations, and speaker variety. To this end, we defined three architectures, inspired from literature baselines. Our results indicate that embedding-based audio encodings have an edge over feature-based encodings. They also suggest that there is a significant margin for improvement, hence the need for different architectures to enable a tighter mutual interaction between input modalities. We speculate that current trends in MMDL, in particular attention-based methods for multimodal input fusion, should be investigated in this domain. We hope that our dataset will facilitate such endeavor. A remarkable result is the performance of the AO configuration, which in some cases is observed to be competitive with TA. This could indicate that, independently of automated speech recognition and transcription systems that may or may not be available for different languages, useful AM systems could be devised to work only based on the audio signal. Possible applications include systems to support debate summarization and news reporting. Future research directions include a more extensive exploration of the possible architectural configurations and embedding methods, and the introduction of attention-based architectural innovations.

Acknowledgements

This work has been partially funded by the European Union's Horizon 2020 Research and Innovation programme under grant agreement 101017142 (StairwAI: Stairway to AI).

References

Yoshua Bengio. 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer.

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.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 5427–5433.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California. Association for Computational Linguistics.

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.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *ACL (1)*, pages 2225–2235. Association for Computational Linguistics.

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.

Sahar Harati, Andrea Crowell, Helen S. Mayberg, and Shamim Nemati. 2018. Depression severity classification from speech emotion. In *EMBC*, pages 5763–5766. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Daniel Kopev, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Detecting deception in political debates using acoustic and textual features. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 652–659.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Zheng Lian, Jianhua Tao, Bin Liu, and Jian Huang. 2019. Conversational emotion analysis via attention mechanisms. In *INTERSPEECH*, pages 1936–1940. ISCA.

Marco Lippi and Paolo Torroni. 2016a. 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.

Marco Lippi and Paolo Torroni. 2016b. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

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.

Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. 2018a. A recorded debating dataset. In *LREC*. European Language Resources Association (ELRA).

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2018b. Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724, Brussels, Belgium. Association for Computational Linguistics.

Habibeh Naderi, Behrouz Haji Soleimani, Sheri Rempel, Stan Matwin, and Rudolf Uher. 2019. Multimodal deep learning for mental disorders prediction from audio speech samples. *CoRR*, abs/1909.01067.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–387, Cham. Springer International Publishing.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Claire Polo, Kristine Lund, Christian Plantin, and Gerald P. Niccolai. 2016. Group emotions: the social and cognitive functions of emotions in argumentation. *Int. J. Comput. Support. Collab. Learn.*, 11(2):123–156.

Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2022. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*, pages 3465–3469. ISCA.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ermal Toto, M. L. Tlachac, and Elke A. Rundensteiner. 2021. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *CIKM*, pages 4145–4154. ACM.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL (1)*, pages 6558–6569. Association for Computational Linguistics.

David Vilares and Yulan He. 2017. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1582, Copenhagen, Denmark. Association for Computational Linguistics.

Serena Villata, Elena Cabrio, Imène Jraidi, M. Sahbi Benlamine, Maher Chaouachi, Claude Frasson, and Fabien Gandon. 2017. Emotions and personality traits in argumentation: An empirical evaluation. *Argument Comput.*, 8(1):61–87.

Yiming Zhang, Ying Weng, and Jonathan Lund. 2022. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12(2).

A Dataset Pre-Processing Details

In this section, we provide information on the debates that were removed owing to issues with the audio file's quality or discrepancy between the audio content and the corresponding transcript. We removed the samples corresponding to the first parliamentary debate in 1988 (Bush vs Dukakis) since the transcript is incomplete and this would have caused alignment mismatches. Regarding the two presidential debates of 2016 (Clinton vs Trump), there was no correspondence between the audio content and corresponding transcripts. Thus, we removed these debates from the original dataset.

The transcript of the first Clinton-Bush-Perot debate of 1992 has been divided into two sections by the Commission. However, the second section did not match the audio file and, thus, we removed the samples corresponding to the second section from the dataset. In the first Carter-Ford debate in 1976, the audio contains a cut of about 30 minutes. Thus, we trimmed the audio file and kept only the audio content before the cut.

B Experimental Setup Details

Table 9 reports the number of samples for each cross-validation fold splits regarding the UKDebates corpus. Likewise, Table 10 provides training statistics for the M-Arg corpus. Table 11 reports the number of samples for the training, validation and test splits of MM-USElecDeb60To16. We used the following seeds for the repeated cross-validation routine: 15371, 15372, 15373. Lastly, Table 12, Table 13 and Table 14 report the class distribution for each train, validation and test split for the UKDebates, M-Arg and MM-USElecDeb60to16 corpora, respectively.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	246	247	247	247	247
Validation	62	62	62	62	62
Test	78	77	77	77	77

Table 9: The number of samples for each train, validation and test fold split regarding the UKDebates corpus.

C Model Calibration

In this section, we report the hyper-parameters set used to calibrate each described classification model. We distinguish between input configurations TA, TO, and AO. In particular, the calibration space for input configuration TA is the combination of those regarding input configurations TO and AO. Table 15 reports the hyper-parameter set used to calibrate the BERT model. Similarly, Table 17 and 16 describe the calibration space of the SVM and Bi-LSTM baselines, respectively.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	1563	1563	1563	1564	1564
Validation	391	391	391	391	391
Test	489	489	489	488	488

Table 10: The number of samples for each train, validation and test fold split regarding the M-Arg ($\gamma \geq 0.85$) corpus.

	No. Sentences
Train	12423
Validation	6894
Test	7464

Table 11: The number of samples for each train, validation and test split regarding the MM-USElecDeb60to16 corpus.

D Performance on Validation Splits

Table 18 reports classification performance on the validation set of the UKDebates corpus. Likewise, Table 19 and 20 report classification metrics for M-Arg and MM-USElecDeb60to16 corpora, respectively.

	Fold 1		**Fold 2**		**Fold 3**		**Fold 4**		**Fold 5**	
	claim	not-claim	claim	not-claim	claim	not-claim	claim	not-claim	claim	not-claim
Train	96	150	98	149	98	149	98	149	97	150
Validation	25	37	24	38	24	38	24	38	24	38
Test	31	47	30	47	30	47	30	47	31	46

Table 12: Class distribution for each train, validation and test split regarding the UKDebates corpus.

	Fold 1			**Fold 2**			**Fold 3**			**Fold 4**			**Fold 5**		
	neither	attack	support	neither	attack	support	neither	attack	support	neither	attack	support	neither	attack	support
Train	1460	19	84	1460	19	84	1460	19	84	1460	20	84	1460	19	85
Validation	365	4	22	365	4	22	366	4	21	366	4	21	366	4	21
Test	457	6	26	457	6	26	456	5	27	456	5	27	456	6	26

Table 13: Class distribution for each train, validation and test split regarding the M-Arg corpus.

	ASD		**ACC**	
	arg	not-arg	claim	premise
Train	9456	2967	5029	4427
Validation	5199	1695	2814	2385
Test	5907	1557	3036	2871

Table 14: Class distribution for each train, validation and test split regarding the MM-USElecDeb60to16 corpus.

Hyper-parameter	**Search Space**
Input Configuration TO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Input Configuration AO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
L2 regularization	$[1e^{-2}, 1e^{-3}, 5e^{-3},$ $1e^{-04}, 5e^{-04}]$
Bi-LSTM units	$[64, 100, 128, 256, 512]$
Audio pooling	$[None, [10, 2], [5, 5],$ $[5, 5, 5]$,$[5], [10, 10]]$
CNN filters	$[8, 64]$
CNN kernel size	$[3, 7]$

Table 15: The hyper-parameters search space of the BERT model.

Hyper-parameter	**Search Space**
Input Configuration TO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
L2 regularization	$[1e^{-2}, 1e^{-3}, 5e^{-3},$ $1e^{-04}, 5e^{-04}]$
Bi-LSTM units	$[32, 64, 128]$
Bi-LSTM layers	$[1, 2]$
GloVe embedding	$[50, 100, 200, 300]$
Learning rate	$[1e^{-3}, 1e^{-4}, 2e^{-4}]$
Input Configuration AO	
Input dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
Classification units	$[64, 100, 128, 256, 512]$
Pre-classification dropout	$[0., 0.1, 0.2, 0.3, 0.4, 0.5]$
L2 regularization	$[1e^{-2}, 1e^{-3}, 5e^{-3},$ $1e^{-04}, 5e^{-04}]$
Bi-LSTM units	$[64, 100, 128, 256, 512]$
Bi-LSTM layers	$[1, 2]$
Audio pooling	$[None, [10, 2], [5, 5],$ $[5], [5, 5, 5], [10, 10]]$

Table 16: The hyper-parameters search space of the Bi-LSTM model.

Hyper-parameter	**Search Space**
Kernel	$[rbf, linear]$
γ	$[5e^{-2}, 1e^{-2}, 1e^{-1}, 5e^{-1}, 1.]$
C	$[0.01, 0.1, 1., 10, 100]$

Table 17: The hyper-parameters search space of the SVM model.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	**66.18**	49.86	48.00	51.58	64.45
M-ArgNet	**71.64**	53.07	63.25	55.02	70.52
BiLSTM	**68.80**	52.90	67.88	49.86	68.06
Random	37.78				

Table 18: Average binary F1-score on the validation set of UKDebates. For each row, we report the best results in bold, second best results are underlined instead.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
SVM	13.26	11.54	12.75	13.50	**24.09**
M-ArgNet	23.69	20.35	23.66	13.04	**26.56**
BiLSTM	21.83	11.98	20.34	11.43	**24.62**
Random	2.62				

Table 19: Average macro F1-score on the validation set of M-Arg ($\gamma \geq 0.85$). For each row, we report the best results in bold, second best results are underlined instead.

		Feature-based		*Embedding-based*	
Model	**TO**	**AO**	**TA**	**AO**	**TA**
ASD					
SVM	**68.01**	56.34	56.76	64.40	67.24
M-ArgNet	66.71	56.14	62.30	63.59	**68.53**
BiLSTM	68.71	58.86	69.35	63.01	**69.39**
Random	50.26				
ACC					
SVM	**66.17**	49.28	49.23	57.72	64.34
M-ArgNet	**68.48**	50.43	67.28	58.36	68.01
BiLSTM	67.78	48.27	68.38	58.30	**68.49**
Random	49.43				

Table 20: Average macro F1-score on the validation set of MM-USElecDeb60to16. For each row, we report the best results in bold, second best results are underlined instead.