M-Arg: Multimodal Argument Mining Dataset for Political Debates with Audio and Transcripts

Rafael Mestre[1,2,†], **Razvan Milicin**[2], **Stuart E. Middleton**[2,†], **Matt Ryan**[1,†], **Jiatong Zhu**[2] and **Timothy J. Norman**[2]

[1] School of Economic, Social and Political Science, University of Southampton
[2] School of Electronics and Computer Science, University of Southampton
†Corresponding authors: {R.Mestre,sem03,M.Ryan}@soton.ac.uk

Abstract

Argumentation mining aims at extracting, analysing and modelling people's arguments, but large, high-quality annotated datasets are limited, and no multimodal datasets exist for this task. In this paper, we present M-Arg, a multimodal argument mining dataset with a corpus of US 2020 presidential debates, annotated through crowd-sourced annotations. This dataset allows models to be trained to extract arguments from natural dialogue such as debates using information like the intonation and rhythm of the speaker. Our dataset contains 7 hours of annotated US presidential debates, 6527 utterances and 4104 relation labels, and we report results from different baseline models, namely a text-only model, an audio-only model and multimodal models that extract features from both text and audio. With accuracy reaching 0.86 in multimodal models, we find that audio features provide added value with respect to text-only models.

1 Introduction

Understanding and modelling argumentation is a recent and key challenge in Natural Language Processing (NLP). Most work addressing this task has focused on extracting arguments from argumentative essays (Stab and Gurevych, 2017), social networks like Twitter (Bosc et al., 2016) or online reviews (Cocarascu and Toni, 2018) and discussions (Habernal and Gurevych, 2017), and not much attention has been paid to mining arguments in natural dialogue. The two most common research questions consider how argumentative relations between units (e.g. support or attack) are annotated or how claims and/or premises are identified (Lawrence and Reed, 2019). We offer, to the best of our knowledge, the first multimodal argumentation mining dataset (M-Arg) of political debates annotated for such argumentative relations of support and attack, using crowd-sourcing techniques. Our contributions are: i) to provide a high quality annotated dataset of political debates with audio and time-stamped transcripts for multimodal argumentation mining; ii) to offer benchmark model results for the research community; and iii) a comparative analysis of the value that multi-modal models bring compared to text-only and audio-only models (Section 5).

The dataset is derived from a collection of US 2020 presidential debates. Five debates were used with the principal speakers being Donald Trump, Joe Biden, Mike Pence and Kamala Harris, and a moderator (Table 1). In three of the debates the candidates spoke only with each other and the moderator, while in the remaining two they interacted with the audience in so-called Town Hall events. The lengths of the audio files ranged from approximately 1 hour to 1 hour 35 minutes. The debates were tokenised by sentences or utterances, with 6527 in total. The relationship between pairs of sentences were then classified by crowd-workers as support, attack or neither using the annotation scheme proposed by Carstens and Toni (2015) (Section 3). The crowd-workers were presented with the sentence pair along with a small extract from the debate to provide context. The resulting dataset consists of 4104 pairs of sentences with the argumentative relationship between them classified, along with features such as the trustworthiness of the crowd-workers, the level of agreement between crowd-workers, and their self-confidence scores (Section 4). Prior to giving details of our methodology, the dataset and comparative analysis, we provide a brief review of related research.

2 Related work

Much of the research in argumentation mining has been dedicated to the identification of argumentative discourse units (ADUs) like claims, major claims and premises. For instance, in a first iteration, Stab and Gurevych (2014) annotated 90

Proceedings of The 8th Workshop on Argument Mining, pages 78–88
Punta Cana, Dominican Republic, November 10–11, 2021.

#	Duration	Split	Speakers	Moderator	# sentences	# labels
1	1:35:06	2	Donald Trump Joe Biden	Chris Wallace	1889	1214
2	1:34:15	2	Donald Trump Joe Biden	Kristen Weller	1648	1018
3	1:31:42	7	Joe Biden Audience	George Steph.	838	519
4	1:00:10	4	Donald Trump Audience	Savannah Guthrie	1132	707
5	1:29:37	2	Mike Pence Kamala Harris	Susan Page	1020	646
					6527	4104

Table 1: Description of the five debates used in the dataset. The column "Split" indicates the number of sub-files in which the audio was split.

persuasive essays with 1673 sentences and 1552 argumentative units. Then, they extended their dataset to 402 essays, achieving a total of 7116 sentences and 6089 argumentative components (Stab and Gurevych, 2017). Carstens and Toni (2015) advocate a relation-based approach towards argumentation mining. Instead of separating the issue of identifying argumentative units and their relation, they reconstitute the task as one of classifying the relationship between sentence pairs as support, attack or neither. They argue that this relation depends upon the context of the discussion. We take the same approach. There are, however, few datasets for relation-based argumentation mining (Paul et al., 2020). Carstens and Toni (2015), for example, annotate 854 pairs of sentences for support/attack without identifying the arguments first. Likewise, the DART dataset (Bosc et al., 2016) consists of 4000 tweets, 446 support relations and 112 attack relations, and Stab and Gurevych (2017) annotate 3616 supports and 219 attack relations in their second version of their essay dataset.

While certain tasks in argument mining have been applied in other disciplines, interdisciplinary approaches are important for the impact of these methods to be fully realised. Some research in political science has started to bridge the gap in tasks like identifying emotion rhetoric (Osnabrüge et al., 2021), gender and emotional expression in politics (Boussalis et al., 2021), emotional mining in political campaigns (Greco and Polli, 2020), lexicometrics of Euromanifestos (Jadot and Kelbel, 2017), and, from the AI perspective, *ethos* mining (Duthie and Budzynska, 2018) using Hansard as a dataset. Argument mining in political debate is, however, still largely to be explored, although Visser et al. (2021) provide an annotation of 2016 US presidential debates with argument types. Benoit et al. (2016) have advocated for the use of crowd-sourced text analysis for political science, finding high levels of agreement and reproducibility between crowd-workers and experts. However, in subjective tasks like identifying support/attack relations, lower levels of agreement are expected. For instance, Faulkner (2014) used Amazon Mechanical Turk (AMT) to annotate 8176 sentences with "for", "against" or "neutral", achieving about 66% of neutral cases and a Cohen's κ of 0.70. Al-Khatib et al. (2020) also used AMT to obtain 16429 labels of different types, including 1736 "relation" labels, defined in their case as "positive", "negative" or "no-argument", with $\kappa = 0.51$.

These datasets focus exclusively on text, and, as far as we can tell, there is not much argument mining research using multiple modalities such as both text and audio, particularly focusing on identifying support or attack relations between ADUs. There are, however, some datasets that could be used by the community for this task. For instance, Mirkin et al. (2019) and Orbach et al. (2020b) provide datasets of debate speeches with transcriptions that could help in the extraction of arguments. Likewise, Mirkin et al. (2020) and Orbach et al. (2020a) come closer to argumentation mining research offering datasets of argumentative content and general-purpose rebuttal in speeches. Also, Kopev et al. (2019) use audio and transcripts of political debates to detect deception. Other research explores emotion recognition or sentiment analysis using the IEMOCAP dataset, which contains text, audio and video with emotion annotations (Busso et al., 2008; Cai et al., 2019). Classic NLP models for relation classification have relied on bag of words (BoW) approaches with common classifiers like random forests, support vector machines or naïve Bayes (Carstens and Toni, 2017), although more recently LSTMs and Bi-LSTMs have been used with good results (Cocarascu and Toni, 2018). Some efforts are being devoted to the use of background knowledge or context. For instance, Paul et al. (2020) proposed Bi-LSTM encoders with self-attention, together with commonsense knowledge extraction. The use of both textual and audio features for the identification of argumentative relations, with approaches similar to those used in multimodal emotion recognition, seems to be mostly unexplored.

Argument mining of political debates can be seen as a long conversation text classification problem where context matters. Unlike the well studied

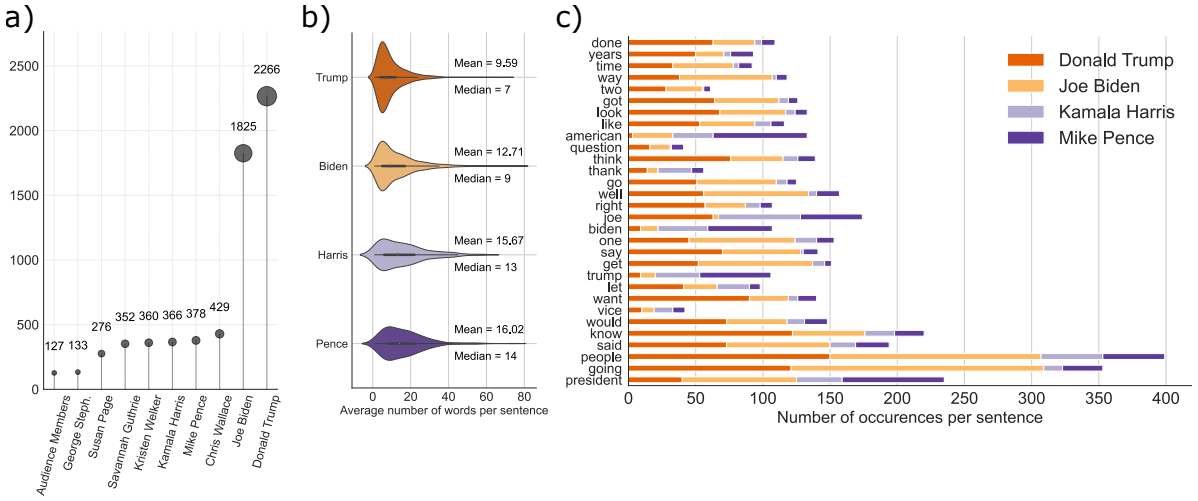

Figure 1: Descriptive visualisation of the original dataset. a) Number of utterances (after sentence tokenisation) by each person across all the debates. Audience members were all aggregated in the label "Audience Members". b) Average number of words per sentence of the four main participants of the debate, showing their density distribution. c) Number of sentences in which the most common used words appeared in by their speaker.

problem area of single and short utterance classification (e.g. 2–3 utterances), dialogue modelling and classification of longer conversations has received little attention to date (Xu et al., 2021). Recent approaches to handle long sequence classification include augmented transformer models with information retrieval (IR) or summarisation models (Xu et al. 2021; Tigunova et al. 2020). We constrain ourselves in this paper to providing results for a set of short conversation classification baseline models as we want to focus on showing the value of using multimodal data. However, we expect recent advances in long conversation classification models to yield good results with our dataset in the future.

3 Methodology

The original source of the M-Arg dataset was available as audio tracks with transcripts from a Kaggle competition.[1] This public-domain dataset was originally constructed by downloading audio from YouTube and transcripts from Rev,[2] as explained in the source metadata. The M-Arg dataset with annotations, full transcripts and audio files, source code and model checkpoints for reproducibility is available online in our GitHub repository.[3]

3.1 Data overview

The original data was presented as audio in .mp3 files and transcriptions in both .txt and .csv files. The .csv files contained three columns: *speaker*, *minute*, and *text*. Since the timestamps did not align perfectly to the audio clips, we performed our own tokenisation and text-audio alignment. The M-Arg dataset associates each sentence with a matched timestamp in the corresponding debate audio file. To do this, each text was split into utterances, defined as single sentences[4]. Visual inspection revealed the transcriptions to be grammatically correct, with no apparent typos and proper use of punctuation, and so automatic sentence-level tokenization performed well. The utterances were then force-aligned to the audio using the web application of the *aeneas* tool[5], obtaining new timestamps. The source audio files were split into different files to comply with the file size limit for the force alignment and to avoid segments where the debate was starting, finishing or going to a break, applause, music, etc. Table 1 summarises the datasets.

Across the five debates, Donald Trump and Joe Biden spoke the most, as can be seen in Figure 1(a). This is expected, as they were both present in three of the five debates. Mike Pence and Kamala Harris, who only participated in one of the debates

[1]The source materials can be found in https://www.kaggle.com/headsortails/us-election-2020-presidential-debates as of August 9th, 2021. The version used was v. 7.

[2]https://www.rev.com

[3]https://github.com/rafamestre/m-arg_multimodal-argumentation-dataset

[4]Using the sentence tokenizer *PunktSentenceTokenizer* from www.nltk.org.

[5]The website of the web application is https://aeneasweb.org/help and their GitHub https://github.com/readbeyond/aeneas/

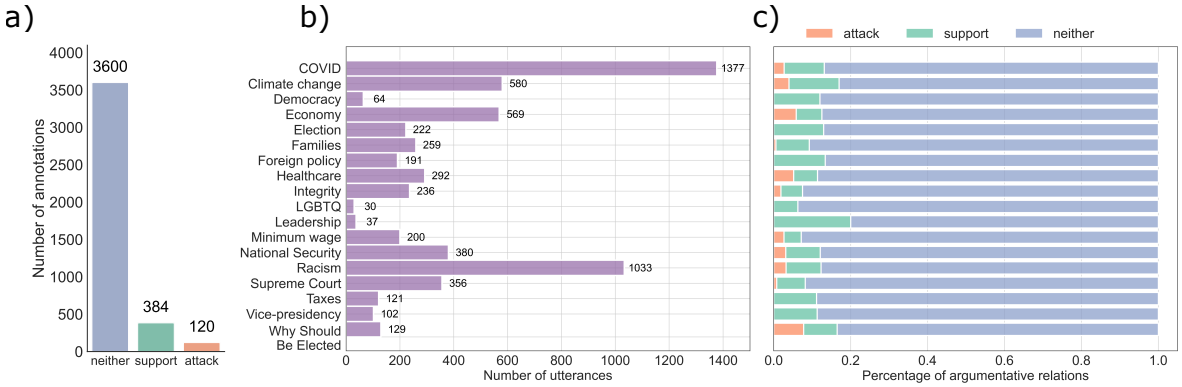

Figure 2: Descriptive visualisation of the annotated dataset, M-Arg. a) Number of pairs annotated for support, attack or neither. b) Total number of sentences in the original dataset labeled as one of the topics in the y-axis. c) Percentage of argumentative relations of pairs of sentences belonging to each one of the topics.

together, spoke a roughly similar number of utterances. Figure 1(b), however, indicates that the main participants in the debate did not speak in a similar manner. The violin plots show the probability density of the average number of words per sentence, and we can observe that Trump spoke sentences of a smaller average length than the rest of the participants. Finally, 1(c) shows the most common words (after removing stop-words) in a stacked barplot according to the speaker (removing moderators and audience members), with certain differences in the usage of words. These results indicate potentially important differences in communication strategies and styles.

3.2 Pairs creation

The M-Arg dataset consists of 4104 labelled pairs of sentences selected from the debates. Sections of the debates were manually labelled by the authors for their "topic", following the explanations of the moderator introducing each section, obtaining high level classifications like "foreign policy". Excerpts of 15 sentences were randomly selected (the "context") and a pair of sentences within the context were chosen to classify their relation (with their distance weighted by a Gaussian distribution to ensure they were close enough). Approximately 1500 sentences were forced to be from different speakers, to balance the dataset by increasing the possibility of finding attack relations. More details on the pair generation strategy and codes can be found in the repository alongside the dataset[6].

[6]GitHub: https://github.com/rafamestre/m-arg_multimodal-argumentation-dataset

3.3 Annotation scheme

The annotation scheme was based on the relation-based argumentation scheme from Carstens and Toni (2015). They argue that an argumentative relation of support or attack is highly dependent on the context. Carstens and Toni (2015) suggest starting from a root claim to construct pairs or match sentences containing the same entities, but we chose to divide them into topics and weight them by distance as explained in Section 3.2. We presented the crowd-workers with a pair of sentences along with the labelled topic of discussion (e.g. "families" or "climate change"), as well as a short 15-sentence extract of the dialogue surrounding these sentences as context. The crowd-workers are asked to use this context, as well as any personal knowledge, to classify the argumentative relation as support, attack or neither, to the best of their ability. By not relying only on the surface meaning of the sentences, we open the way for the use of this dataset in more complex scenarios. For instance, it could be applied together with long- or short-text summarisation to take into account the context in a dialogue (Xu et al., 2021) or knowledge-based models linked to databases or fact-checking websites (Paul et al., 2020). Consider the following pair:

- **Joe Biden:** It's criminal.
- **Donald Trump:** They are so well taken care of.

At a first glance, it is not possible to know what Biden and Trump are talking about. We might assume that the relationship is attack, but this would be a big assumption only based on the fact that they are opposing candidates. Reading the context, we

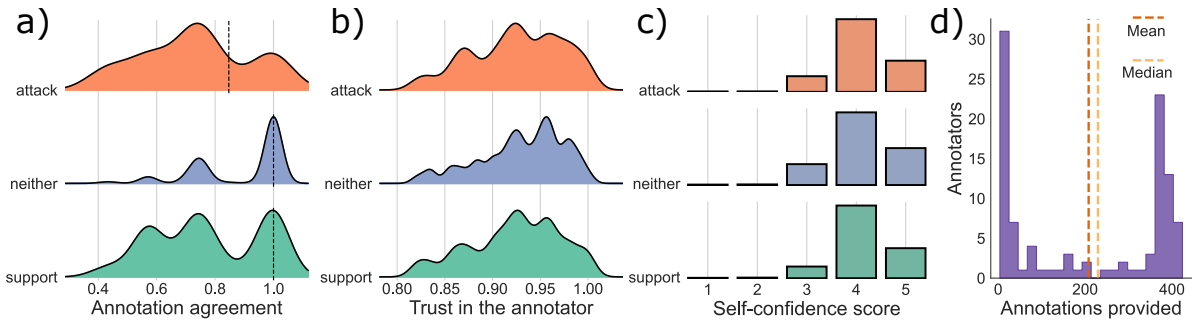

Figure 3: Relationship between annotations and confidence parameters. a) Distribution of annotations according to the annotation agreement; b) to the trust in the annotator; and c) to the self-confidence score given by the annotator. d) Distribution of the annotators with respect to the number of annotations they provided.

might find out that they are actually talking about the infamous controversy of the U.S. Immigration and Customs Enforcement's (ICE) camps, where children were put in cages and separated from their parents. Joe Biden claims that what the Trump administration has done is criminal. Trump answers by saying that they (the children) are "so well taken care of" because reporters went there and saw that the facilities were very clean. The argumentative relation between these two sentences is clearly an attack. Indeed, out of the 81 annotators who classified this test question, 84% of them agreed that this was an attack relation. Work on context summarisation or knowledge extraction could help train models to understand why this was an attack.

In our dataset, ADUs consist of sentences fully delimited by periods, but in many cases they will not (Lawrence and Reed, 2019). They might span several sentences or even be one of the clauses within a sentence. Likewise, during a heated debate, the structure of the argument might not be easily identifiable:

- **Joe Biden:** We learned that this president paid 50 times the tax in China as a secret bank account with China, does business in China, and in fact, is talking about me taking money?
- **Joe Biden:** What are you hiding?

This pair might be interpreted together as a legitimate question being raised to attack the *ethos* of Donald Trump. However, in and of itself, the second sentence can be taken as a rhetorical way of claiming that Trump is hiding something. In this context, the first sentence is supporting this claim. Annotators did well in this task, with 83% of judgements of this test question correctly labelling it as support.

3.4 Crowd-worker annotation

For the annotation task, the platform Appen was used.[7] Crowd-workers were presented with a pair of sentences, topic and context. They were asked to classify the argumentative relation and report their confidence on a Likert scale, ranging from 1: "not confident at all", to 5: "very confident". Each worker was then paid per "page" of work completed, with each page containing between 4 and 6 tasks. To ensure accuracy in the annotations, the contributors were quizzed at the beginning and during the annotations (once per page) with test (or gold) questions. The trust in the annotator was thus defined as the percentage of test questions that they answered correctly, and we set an accuracy threshold of 81%. Other quality settings were enabled, such as: minimum time spent per page to 90 seconds and no more than 60% of supports and 35% of attacks classified. If the annotators did not meet any of these standards, their judgements were not used. Dynamic judgements, that could range from 3 to 7 annotations if the agreement in the annotation was below 70%, were also enabled to improve the agreement of each annotation. A total of 101 test questions were used in this annotation and 104 reliable workers participated, out of 287 that attempted it. Overall, considering the quality settings (e.g. dynamic judgements, tainted answers), 21646 reliable annotations were collected (5746 belonging to gold questions and 15900 to random pairs), and a separate 1663 annotations were rejected.[8]

[7] https://appen.com/.

[8] Extensive details of the annotation scheme and quality settings can be found in the GitHub of the project: https://github.com/rafamestre/m-arg_multimodal-argumentation-dataset.

4 Dataset

4.1 Description and relevant examples

The M-Arg dataset consists of a total of 4104 pairs of sentences (including golden ones), of which 384 are support relations, 120 are attack relations and 3600 are neither support nor attack, as shown in Figure 2(a). Despite efforts to increase the number of support/attack relations, as explained in Section 3.2, the dataset is imbalanced towards the neither side. This is nevertheless expected, as most of the utterances during a debate are not argumentative in nature. Eighteen different topics were identified in the debates, with the most common ones being "COVID", "Racism", "Climate change" and "Economy". Figure 2b shows the total number of utterances from each topic throughout the whole dataset.[9] Some topics, such as "LGBTQ" or "Leadership" had very few instances, since they were only discussed briefly in one of the debates. Many of these topics, however, could be combined, such as "Taxes" and "Economy", as desired. For each topic we can see in Figure 2(c) the distribution of argumentative relations that were annotated. Topics such as "Foreign Policy" or "Taxes" did not contain attack relations, most likely due to the fact that those sections were small.

Whether an argument is supporting or attacking a claim is a subjective matter. Philosophy of argumentation has attempted to establish more or less general argumentation frameworks with different categorisations. However, it is almost certain that thresholds for what quality of information supports or attacks an argument, or judgements on whether such argument is sufficiently valid or not vary by person and context. Our annotated dataset, thus, provides a collective representation of how people reason and understand arguments, and a large number of disagreements are expected. Indeed, the prevalence of fake news or fallacies, or even reasonable disagreements over interpretation of values and inferences in everyday political discourse, has shown us that the same premises can be deemed supports or attacks in different contexts. As we cannot expect that everyone thinks of arguments or fallacies in the same way, the annotation task needs to be accessible and understandable but still closely guided and validated by the theoretical frameworks to reflect informed but real interpretations of support and attack in open dialogue in political domains. In the instructions for the annotation task, the contributors were asked to focus on whether a sentence provided a reason that supported or attacked its counterpart, in order to avoid confusing an attack *towards something/someone* with an attack *towards a claim*. Consider the following example in which Biden is not providing any reason to support his claim, although it was meant to attack Trump. Many people would interpret this as attack (as many annotators did):

- **Donald Trump:** There's abuse, tremendous abuse.
- **Joe Biden:** Simply not true.

This case was correctly labeled as neither, but only with an agreement score of 56%. In other cases, support was simply confused with repetition:

- **Joe Biden** And what's happening is too many transgender women of color are being murdered.
- **Joe Biden:** They're being murdered.

This was annotated as support by three crowd-workers, but this is simply coherence between the sentences or a simple reiteration of a claim. This subjectivity is summarised in Figure 3(a)-(c), which shows the agreement score, the trust in the annotator and the self-confidence score for each label. In general, crowd-workers labeled relations independently of their trust score and their self-confidence score. However, attack relations were more controversial with 25% of annotations above 0.87, whereas for support and neither at least 25% of annotations had an agreement of 1. Overall, the average agreement was 0.87 and the median agreement 1.

5 Evaluation

5.1 Crowd-worker agreement

The presence of subjectivity leads us to evaluate the agreement among crowd-workers (also known as intern-annotator agreement) using Krippendorf's α (Krippendorff, 1980). This agreement score allows for a variable number of annotations in each instance, with an unspecified number of crowd-workers that do not necessarily need to annotate every single instance, making it suitable for our case. Considering all the annotations, we obtained

[9] Due to the constraints in random pair generation, the topic distribution in the annotated dataset differs slightly, but the distribution closely resembles that of the original dataset.

	All annotations					Self-confidence = 5				
Trust	α	**Workers**	**# Annots.**	**Supports**	**Attacks**	α	**Workers**	**# Annots.**	**Supports**	**Attacks**
≥ 0.80	**0.43**	104	21646	4370	2036	0.57	93	5941	1133	530
≥ 0.85	0.44	96	20342	4056	1919	0.59	88	5666	1069	514
≥ 0.90	0.46	76	16747	3251	1508	0.63	69	4615	792	407
≥ 0.95	0.53	53	9066	1562	776	0.72	46	2393	346	175
$= 1$	0.44	27	746	248	117	0.79	25	227	74	31

Table 2: Krippendorff's α values for different filterings of the data. Notice the value in bold corresponds to the overall α from the whole dataset, since our trust threshold was ≥ 0.81. With decreasing number of annotations, high fluctuations in α are to be expected, hence the smaller value of 0.44 for the highest trust.

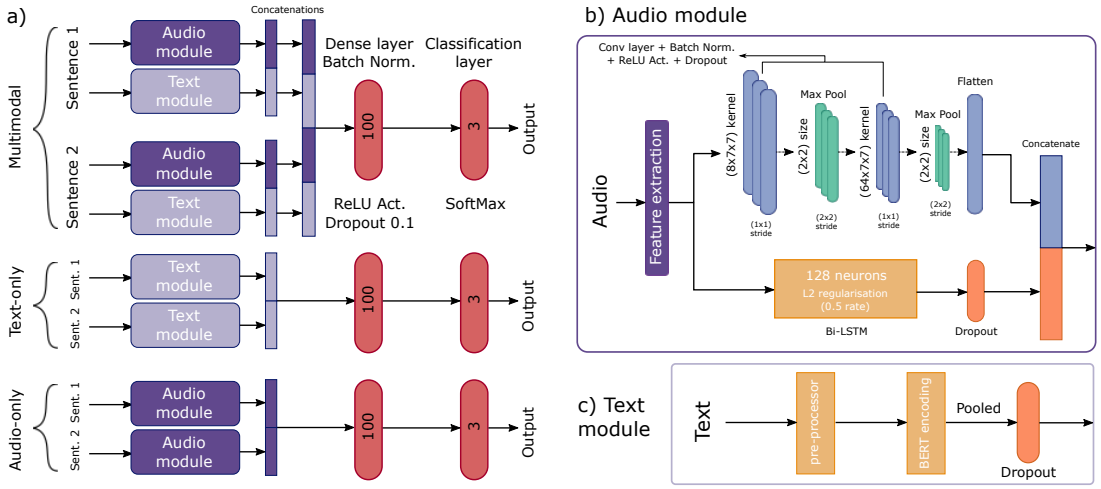

Figure 4: Schematic of the relation classification models. a) In the multimodal model each sentences is passed in parallel through an audio module and a text module. b) The text module consists only on a BERT encoding layer with dropout. c) The audio module is based on parallel CNN and Bi-LSTM.

$\alpha = 0.43$. Considering that the distribution of annotations show that some contributors annotated many sentences, while others very few (Figure 3(d)), we filtered by the most diligent workers but found no significant change in α.[10] However, given that annotators are assigned a trust score and they annotated with different self-confidence, we calculated different α's by filtering by these values. Table 2 shows the α scores when we filter by crowd-worker trust rating for all annotations (left) and only for those that were annotated with the maximum confidence (right). We can see that the crowd-worker agreement increases up to 0.53 when we filter annotators with higher trust (although it drops again for the maximum trust), however at the cost of decreasing the number of workers and annotations in the dataset. Likewise, if we only consider those annotations that were provided with high certainty, we see an overall large value of 0.57, going up to 0.79 if we also filter for high trust workers. Other reports in the literature classifying argumentative relations have yielded Cohen's $\kappa = 0.70$ (Faulkner, 2014), $\kappa = 0.51$ (Al-Khatib et al., 2020), $\alpha = 0.67$ (Bosc et al., 2016), $\alpha = 0.81$ (Stab and Gurevych, 2014).

It might seem surprising to obtain a low crowd-worker agreement, given that the average agreement in the annotation was 85%, as mentioned above. These numbers, however, need to be considered with care. Krippendorff's α measures disagreement beyond that expected by chance, but our data is not balanced, so our labels are not equally probable. It has been observed that Krippendorff's α can be heavily attenuated in imbalanced datasets (Jeni. et al., 2013). Indeed, if we sub-sample our dataset for 100 attack, support and neither relations, we find $\alpha = 0.540 \pm 0.015$ (standard deviation after 10 trials). If we sub-sample it unbalanced, with 10 attack, 10 support and 1000 neither, we find $\alpha = 0.170 \pm 0.047$ (standard deviation after 10 trials). This is a big difference in values, even though

[10] Krippendorff's α was calculated with the *nltk.metrics.agreement* module (v. 3.6). Results were double-checked with the *krippendorff* module, https://github.com/pln-fing-udelar/fast-krippendorff, which yielded almost identical results.

Model	Text dropout	Audio dropout	Acc.	M-F_1	w-F_1	Attack m-F_1	Neither m-F_1	Support m-F_1
Audio only		0.2	0.84	0.41	0.81	0.18	0.91	0.14
Text only	0.1		**0.86**	0.37	0.82	0.06	**0.92**	0.14
Multimodal	0.1	0.1	**0.86**	0.41	**0.83**	0.16	**0.92**	0.14
		0.2	0.85	**0.45**	**0.83**	**0.24**	**0.92**	**0.21**
Multimodal ≥0.85 agreement	0.1	0.2	**0.91**	0.40	**0.90**	0.12	**0.95**	0.10

Table 3: Models' performance. M-F_1 stands for macro-averaged F_1, w-F_1 for weighted-F_1, and m-F_1 for micro-averaged F_1.

the source data is the same. In any case, given the subjectivity of the task, we do not believe a small α to be necessarily a bad result, since many judgements might not lend an obvious collective answer and, most importantly, people might believe one instance is a supportive argument, while others believe it is not an argument at all. We believe there is significant value in these unclear annotations, as they give insight into how people understand the arguments put forward in political debate.

5.2 Argumentative relation classification

To measure the quality of our corpus and study the potential added value of audio features in argumentation mining, we evaluated the performance of different classification models based on a multimodal model, as well as text-only and audio-only models (Figure 4(a)). First, the input pair of sentences were split into audio and text. In the multimodal model, each sentence pair was passed through an audio and text module and their outputs concatenated, passed through a 100-unit middle layer and a 3-output classification layer. In the text-only and audio-only models, the sentences were only passed through the text or audio module, respectively, and the middle and classification layers were the same. The audio module shown in Figure 4(b) was based on a previous model by Cai et al. (2019) for multimodal emotion recognition and consisted of a feature extraction module followed by a CNN in parallel with a Bi-LSTM, chosen to maximise the extraction of local and global features. The text-only module Figure 4(c) consisted of a BERT preprocessor and a BERT encoding of L=12 hidden layers (i.e., Transformer blocks), a hidden size of H=768, and A=12 attention heads. The missing dropout rates can be found in Table 3.

Audio feature extraction was performed using the Python module "librosa" (McFee et al., 2015). The features were: Mel-frequency cepstral coefficients (MFCCs), which are widely used features for characterising and detecting voice signals (Klapuri and Davy, 2006); several spectral features like spectral centroids (Klapuri and Davy, 2006), spectral bandwidth (Klapuri and Davy, 2006), spectral roll-off (McFee et al., 2015) and spectral contrast (Jiang et al., 2002); and a 12-bit chroma vector (McFee et al., 2015). For each sentence, we used the timestamp to clip the audio file with a buffer of ± 2 s to ensure the full audio of the utterance was captured.

To train the three models, we used the Adam optimiser with a learning rate of 0.00005, a batch size of 16 and 50 epochs. A time-based learning rate schedule function was used with a decay rate of 0.0000002 and the loss function was categorical cross-entropy. Table 3 shows the evaluation metrics of these models. The values in "text dropout" and "audio dropout" are the rates of all the dropout layers of the respective models (Figure 4). All three models perform well identifying neither labels, but they do not perform so well identifying attacks or supports, with the highest F_1 values being 0.24 and 0.21, respectively. The text-only model fails to identify attack relations to the same level of the audio-only and multimodal models, most likely due to the imbalance of the data. A multimodal model with a dropout rate of 0.2, however, increases the F_1 for attacks from 0.06 in the text-only model to 0.24, and for supports from 0.14 to 0.21. Surprisingly, the audio-only model performs better than the text-only model in identifying attacks and neither, and closely matches the multimodal model. As proof of concept, we filtered the annotations by their agreement ($\geq$0.85) and we assessed the best performing multimodal model. We obtained an even higher accuracy value of 0.91, especially for neither labels, with m-F1 of 0.95, although identification of support and attack relations was worse, most likely due to a decrease of useful labels. Over-

all, we believe that audio provides relevant features for the identification of argumentative relations and its added value with respect to text helps recapitulate the complexity of this type of data in heavily unbalanced datasets.

6 Conclusions and future work

In this paper, we have presented a multimodal argumentation mining dataset (M-Arg) for political debates based on a corpus of the US 2020 presidential debates with audio and transcripts. The dataset was annotated using crowd-sourcing techniques and we present descriptive statistics of the dataset itself, as well as of the annotations, with discussion of some interesting examples. As a baseline for future research, we evaluated the classification performance of audio-only, text-only and multimodal models. We found that the audio-only and multimodal models could perform with high levels of accuracy and F_1, although they encountered problems classifying support and attack relations very efficiently. The text-only model performed similarly, but its accuracy in attack classifications was low due to the imbalance of the data. Adding the audio to the text, however, in a multimodal model, helped increase the metrics of both support and attacks, although they still remained quite low. We believe these to be encouraging results, as improvements like reinforcement learning to tackle data imbalance, optimised extraction of audio features, addition of (cross-)attention layers, summarisation of the surrounding context or use of background knowledge databases, could greatly improve these performance metrics. Moreover, the data can be filtered according to annotation agreement, the annotator's trust and self-confidence in the annotation, potentially training models with higher precision and/or recall, although with less data.

One limitation of our dataset is that ADUs are defined in a very simple manner (by a period with tokeniser). ADUs might be full sentences on certain occasions, but they might encompass several sentences or simply a clause within one. Further work to improve this dataset would include the identification of ADUs (without necessarily labelling them as claim or premise). Likewise, even if a sentence contains a full ADU, in natural dialogue it might not present as a clearly stated claim or premise, but might contain irony or rhetorical questions.

As already discussed, whether a pair of sentences are showing support or attack is a somewhat subjective matter, and for that reason we obtain Krippendorff's $\alpha = 0.43$. One cannot expect crowdworkers to identify, or even easily understand, all types of arguments or what constitutes a fallacy; philosophers continue to disagree. Yet with some instruction and information these annotations can better reflect real-world judgements about support and attack arguments. For certain applications, especially where including marginalised voices, AI systems will need to understand and detect how people argue, even if they do not follow the dictates of argumentation theory (Young, 2000). We believe our dataset will be of interest for understanding what people think a proper argument is.

Acknowledgements

This work has been funded by the Web Science Institute of the University of Southampton. The authors would also like to acknowledge the support of UK Research and Innovation (UKRI) funding (grant ref MR/S032711/1).

Ethical considerations

Ethics approval for this research was received from the University of Southampton's Faculty of Social Science Ethics and Research Governance committee, Ref: 66226, Date 22/07/2021. The original dataset in which we build the M-Arg dataset was available under license CC0: Public Domain. Fair treatment of the workers involved in the annotation of the dataset was ensured by Appen's code of ethics (`https://appen.com/crowd-wellness/`). We aimed at providing a fair wage for the work provided and, according to the platform statistics, the workers received an hourly compensation with median $7.69 and interquartile mean $7.57. Before sharing our annotated dataset, we have stripped all information that could be potentially sensitive, such as IP's or locations, and we have re-anonymised the anonymous worker ID's that were provided.

References

K. Al-Khatib, Y. Hou, H. Wachsmuth, C. Jochim, F. Bonin, and B. Stein. 2020. End-to-end argumentation knowledge graph construction. *Proc. AAAI Conference on Artificial Intelligence*, pages 7367–7374.

K. Benoit, D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov. 2016. Crowd-sourced text analysis: Reproducible and agile production of po-

litical data. *American Political Science Review*, 110(2):278–295.

T. Bosc, E. Cabrio, and S. Villata. 2016. DART: A dataset of arguments and their relations on twitter. *Proc. 10th Int. Conf. on Language Resources and Evaluation*, pages 1258–1263.

C. Boussalis, T. G. Coan, M. R. Holman, and S. Müller. 2021. Gender, candidate emotional expression, and voter reactions during televised debates. *American Political Science Review*, pages 1–16.

C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

L. Cai, Y. Hu, J. Dong, and S. Zhou. 2019. Audio-textual emotion recognition based on improved neural networks. *Mathematical Problems in Engineering*, 2019.

L. Carstens and F. Toni. 2015. Towards relation based argumentation mining. In *Proc. 2nd Workshop on Argumentation Mining*, pages 29–34.

L. Carstens and F. Toni. 2017. Using argumentation to improve classification in natural language problems. *ACM Trans. on Internet Technology*, 17(3):1–23.

O. Cocarascu and F. Toni. 2018. Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics*, 44(4):833–858.

R. Duthie and K. Budzynska. 2018. A deep modular RNN approach for ethos mining. *Proc. Int. Joint Conf. on Artificial Intelligence*, pages 4041–4047.

A. R. Faulkner. 2014. *Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization*. Ph.D. thesis, CUNY.

F. Greco and A. Polli. 2020. The political debate on immigration in the election campaigns in Europe. In *Springer Proceedings in Complexity*, pages 111–123. Springer, Cham.

I. Habernal and I. Gurevych. 2017. Argumentation mining in user-generated Web discourse. *Cmputational Linguistics*, 43(1):125–179.

C. Jadot and C. Kelbel. 2017. 'Same, same, but different.' Assessing the politicisation of the European debate using a lexicometric study of the 2014 Euro-manifestos. *Politique Europeenne*, 55(1):60–85.

L. A. Jeni., J. F. Cohn, and F. De La Torre. 2013. Facing imbalanced data - Recommendations for the use of performance metrics. *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 245–251.

D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai. 2002. Music type classification by spectral contrast feature. *Proc. IEEE Int. Conf. on Multimedia and Expo*, pages 113–116.

A. Klapuri and M. Davy. 2006. Signal processing methods for music transcription. In *Signal Processing Methods for Music Transcription*, chapter 5. Springer Science & Business Media.

D. Kopev, A. Ali, I. Koychev, and P. Nakov. 2019. Detecting deception in political debates using acoustic and textual features. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pages 652–659.

K. Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology.* SAGE Publications, Inc, Thousand Oaks, CA.

J. Lawrence and C. Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto. 2015. librosa: Audio and music signal analysis in Python. *Proc. 14th Python in Science Conference*, pages 18–24.

S. Mirkin, M. Jacovi, T. Lavee, H. K. Kuo, S. Thomas, L. Sager, L. Kotlerman, E. Venezian, and N. Slonim. 2019. A recorded debating dataset. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 250–254.

S. Mirkin, G. Moshkowich, M. Orbach, L. Kotlerman, Y. Kantor, T. Lavee, M. Jacovi, Y. Bilu, R. Aharonov, and N. Slonim. 2020. Listening comprehension over argumentative content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 719–724.

M. Orbach, Y. Bilu, A. Gera, Y. Kantor, L. Dankin, T. Lavee, L. Kotlerman, S. Mirkin, M. Jacovi, R. Aharonov, and N. Slonim. 2020a. A dataset of general-purpose rebuttal. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5591–5601. Association for Computational Linguistics.

M. Orbach, Y. Bilu, A. Toledo, D. Lahav, M. Jacovi, R. Aharonov, and N. Slonim. 2020b. Out of the Echo Chamber: Detecting Countering Debate Speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7073–7086.

M. Osnabrüge, S. B. Hobolt, and T. Rodon. 2021. Playing to the gallery: Emotive rhetoric in parliaments. *American Political Science Review*, pages 1–15.

D. Paul, J. Opitz, M. Becker, J. Kobbe, G. Hirst, and A. Frank. 2020. Argumentative relation classification with background knowledge. *Frontiers in Artificial Intelligence and Applications*, 326:319–330.

C. Stab and I. Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proc. 25th Int. Conf. on Computational Linguistics*, pages 1501–1510.

C. Stab and I. Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

A. Tigunova, A. Yates, P. Mirza, and G. Weikum. 2020. CHARM: Inferring personal attributes from conversations. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5391–5404.

J. Visser, J. Lawrence, C. Reed, J. Wagemans, and D. Walton. 2021. Annotating argument schemes. *Argumentation*, 35(1):101–139.

J. Xu, A. Szlam, and J. Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. arXiv:2107.07567.

I. M. Young. 2000. *Inclusion and Democracy*. Oxford University Press.