Hitting your MARQ: Multimodal ARgument Quality Assessment in Long Debate Video

Md Kamrul Hasan[1], James Spann[1], Masum Hasan[1], Md Saiful Islam[1], Kurtis Haut[1], Rada Mihalcea [2], Ehsan Hoque[1]

1 - Department of Computer Science, University of Rochester, USA

2 - Computer Science & Engineering, University of Michigan, USA

{mhasan8, jspann2}@cs.rochester.edu, {mhasan, mislam6, khaut}@ur.rochester.edu

mihalcea@umich.edu, mehoque@cs.rochester.edu

Abstract

The combination of gestures, intonations, and textual content plays a key role in argument delivery. However, the current literature mostly considers textual content while assessing the quality of an argument, and is limited to datasets containing short sequences (18-48 words). In this paper, we study argument quality assessment in a multimodal context, and experiment on DBATES, a publicly available dataset of long debate videos. First, we propose a set of interpretable debate-centric features such as clarity, content variation, body movement cues, and pauses, inspired by theories of argumentation quality. Second, we design the **M**ultimodal **AR**gument **Q**uality assessor (MARQ) – a hierarchical neural network model that summarizes the multimodal signals on long sequences and enriches the multimodal embedding with debate-centric features. Our proposed MARQ model achieves an accuracy of 81.91% on the argument quality prediction task and outperforms established baseline models with an error rate reduction of 22.7%. Through ablation studies, we demonstrate the importance of multimodal cues in modeling argument quality.

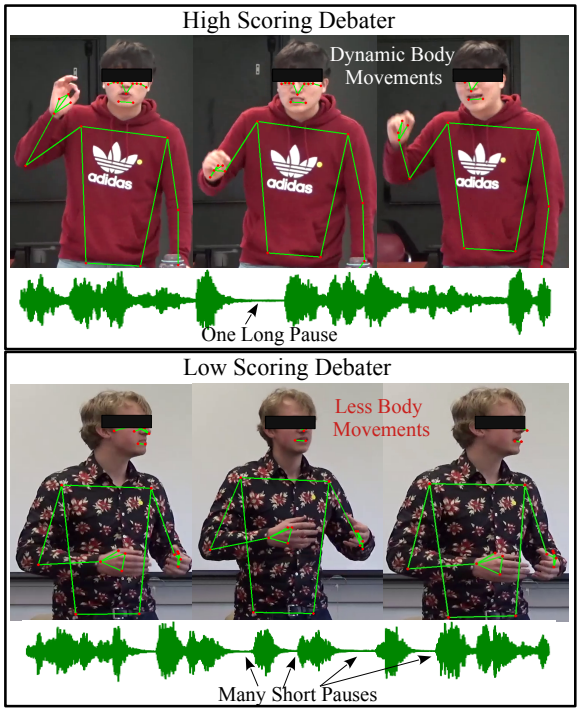

Figure 1: Delivering an argument involves multiple modalities. For example, a high scoring debater may use a certain prosodic style and dynamic hand gestures to present an argument more effectively, whereas a low scorer may hesitate (i.e., frequent short pauses) or show lack of confidence. Language only analysis fails to understand these extra cues.

1 Introduction

Structured debates and discussions are the basis for expressing opposing opinions, and are a tool for convincing others to share that opinion. Starting with a topic to argue, one can outline steps to reach a conclusion of why that topic is correct. This can take many forms in day-to-day life ranging from salesmen upselling a product or presidential debates, to people arguing whether to get vaccinated or to wear a mask.

While the points of the argument may be valid, certain attributes such as clarity in the text, hand movements, and spoken style increase the effectiveness of the argument (Wachsmuth et al., 2017a; Braga and Marques, 2004; Straßmann et al., 2016). These attributes increase the credibility of the speaker, and their ability to convince the listener (Figure 1). Measuring the quality of an argument given the language and other non-verbal features remains an elusive problem. Although argument quality assessment is an established research area in NLP, assessment in a multimodal context is understudied. Most of the previous work focused on argument quality prediction on *short* text sequences (22-48 words). However, often longer text sequences are needed to validate an argument on a certain topic. Even when a model like BERT (Devlin et al., 2018) can be trained to associate raw text with a quality metric (Gretz et al., 2020;

Toledo et al., 2019), it can be difficult to interpret which features lead to the output score. Such kind of interpretability is crucial to design a feedback system for people who want to improve their communication skills (Fung et al., 2015).

In this paper, we study argument quality assessment in a multimodal context using DBATES (Sen et al., 2021) – the largest (N=716) publicly available dataset of debate videos. We design interpretable debate-centric features (DCF) such as content variation, clarity, pauses, hand movement, emotional appeal, and so on based on theories of argument quality (Wachsmuth et al., 2017a; Braga and Marques, 2004; Straßmann et al., 2016). Moreover, we propose a hierarchical multimodal model named MARQ (**M**ultimodal **AR**gument **Q**uality assessor) to predict high vs. low quality arguments in *long* debate speeches (6 minutes recordings & 1500 words). Sentence level rich unimodal embeddings are extracted from pretrained models (e.g Universal Sentence Encoder (Cer et al., 2018), Wav2Vec2 (Baevski et al., 2020)) to reduce long sequential dependency. A set of LSTM encoders and a Multihead Self-Attention layer are used to capture the interaction across the intra-modal, inter-modal, and DCF information. Our main contributions are:

- We present the first comprehensive study on multimodal argument quality assessment. A set of interpretable debate-centric features are derived based on the theories of argumentation quality. These features are statistically significant and can achieve 75.53% accuracy in argument quality prediction task, thus validating their usefulness.

- We propose MARQ – a hierarchical multimodal model that captures the long-term dependencies and the complex interactions among the modalities. The model achieves 81.91% accuracy in distinguishing the quality of arguments and outperforms several established baselines with an error rate reduction of 22.7%.

2 Dataset

DBATES (Sen et al., 2021) is a publicly available dataset collected from the 2019 North American Debating Championship. The tournament followed the British parliamentary debate format for university-level (Eckstein and Bartanen, 2015). A motion (e.g., "parents should teach their children that they are inherently special") is given 20 minutes before each debate. Eight debaters are split into two parties – Government and Opposition. The Government party present arguments to support the given motion while the Opposition party argues against the motion. Each debater gets 6 minutes to present their arguments to support their stance. Expert judges discuss among themselves and assign a score (within 50-100) to each person's performance based on the quality of the argument. A total of 716 debate videos (6 minutes each) from 140 unique debaters have been recorded. The median score (77) is used as a threshold to distinguish between high and low-quality arguments. During the final rounds of the debate championship, the judges have provided the list of winners instead of assigning scores to each debate speeches. We remove these instances (79 samples) and use the debate speeches that have been annotated with the score (within 50-100).

Dataset	N	Modalities	Mean Seq Len
IBM-Rank-Args	6.3k	{*l*}	22.98 words
IBM-Rank-Pairs	14k	{*l*}	23.47 words
IBM-Rank-30k	30K	{*l*}	18.22 words
UKPConvArg1	11k	{*l*}	48.3 words
DTC	400	{*l*,*a*,*v*}	22.5 seconds
DBATES	716	{*l*,*a*,*v*}	**6 minutes, 1500 words**

Table 1: Comparison between DBATES and other datasets. Here *l, a, v* represent language (text), audio, and visual modalities respectively. DBATES presents a unique challenge due to the sequence length being significantly longer (at least 13x) than the previous ones and having all three modalities.

Table 1 presents a comprehensive comparison among the existing datasets. Most of the existing research is limited to datasets (e.g., IBM-RANK (Gretz et al., 2020; Toledo et al., 2019) and UKP-ConvArg1 (Habernal and Gurevych, 2016a)) containing only language (text) and with smaller sequences (the average sequence length is 18-48 words). These arguments are collected and annotated through crowd-sourcing. Debate Trainees Corpus (DTC) Petukhova et al. (2017) is a multimodal debate dataset consisting of 400 arguments totaling 2.5 hours. However, the dataset is not publicly available.

We choose DBATES to study multimodal argument quality analysis, since it is the largest and only multimodal debate dataset that is publicly available for research. Moreover, the dataset is collected from a competitive college debate competition and has been annotated by expert judges. The aver-

age sequence length is around 1500 words which is significantly longer compared to other datasets. The long multimodal sequence is particularly challenging for neural models to comprehend, which is applicable to other multimodal tasks as well.

3 Debate Centric Features

Argument quality can be assessed in many different granularity, some of them are subjective and difficult to compute. Here, we propose a set of computable and objective debate-centric features considering all of the language, acoustic and visual modalities. Experiments in later sections show that these features can discriminate between high and low-quality arguments.

3.1 Language-DCF

Content Variation: Monotonous speech that involves repetition and less diversity in content can reduce the effectiveness of the argument. We assume that, as a whole, a segment of sentences discuss the central topic. If all the sentences of that segment are very similar to the central topic, i.e., there are less variation in content, the argument may become repetitive or monotonous. Each debate consists of multiple segments like introduction, constructive, rebuttal, conclusion and so on. In order to measure variation in content, we first use an Universal Sentence Encoder (USC) (Cer et al., 2018) embedding of the whole segment to represent the central topic, and USC embedding of each sentence within the segment to represent the local topic. The average cosine distance between a sentence embedding and the corresponding segment embedding provides an approximation of the content variation present in the argument.

Emotional Appeal: Emotional appeal makes the target audience more receptive to the stance of the speaker's argument (Wachsmuth et al., 2017a). To represent emotional appeal, we compute the sentiment (positive/negative) and emotion (sadness, joy, fear, disgust, anger) scores of each sentence using IBM Bluemix (Gheith et al., 2016).

Clarity: A clear argument that can avoid ambiguity and unnecessary complexities can easily persuade the target audience (Wachsmuth et al., 2017a). We extract Flesch Reading Ease metric (Flesch and Gould, 1949) to measure the clarity of an argument. The metric assigns a readability score (between 0 and 100) to a given text, high score indicating the text is easy to understand. Sentence structure complexity also affects the clarity of the text. We also extract fourteen features that represents the syntactic complexity of a sentence (Lu, 2010). The features are mean length of sentence (MLS), mean length of T-unit (Hunt, 1965) (MLT), mean length of clause (MLC), clauses per sentence (C/S), verb phrases per T-unit (VP/T), clauses per T-unit (C/T), dependent clauses per clause (DC/C), dependent clauses per T-unit (DC/T), T-units per sentence (T/S), complex T-unit ratio (CT/T), coordinate phrases per T-unit (CP/T), coordinate phrases per clause (CP/C), complex nominals per T-unit (CN/T), and complex nominals per clause (CP/C).

LIWC Features: LIWC features consist of word counts for each of the 80 semantic classes present in the LIWC lexicon (Pennebaker et al., 2001). Some categories include the frequency of concessive subordinates (e.g., although, though); conjuncts (e.g., alternatively, on the other hand); negations (e.g., no, neither, nor) and causal conjuncts (e.g., consequently, therefore) which are often used in a argument to present the logic.

3.2 Acoustic-DCF

The prosodic style can play a key role while delivering an argument. Variation of pitch, showing control on the pauses and speed of speech, and a smooth delivery can be perceived as expressions of enthusiasm, engagement, commitment and charisma (Rosenberg and Hirschberg, 2009), which helps to persuade the audience to make the argument more credible. On the contrary, taking frequent pauses and unclear articulation can hurt the effectiveness of an argument delivery. These are applicable even if the textual content remains the same, implying the language only assessment of argument quality will fail to consider these factors. We use Opensmile (Eyben et al., 2010) to capture pitch, and commonly used variants of jitter and shimmer. We model pause as one second silence in the audio, and extract both the number of pauses and their duration as Acoustic-DCF features.

3.3 Visual-DCF

Body language plays an important role to show the confidence in a speaker and increase the credibility of the argument to the audience. Moving the arms, stemming the hands on the hip increase the dominance perception of the speaker (Straßmann et al., 2016). We extract upper body landmarks from each frame using Mediapipe[1] (Bazarevsky

[1]https://mediapipe.dev/

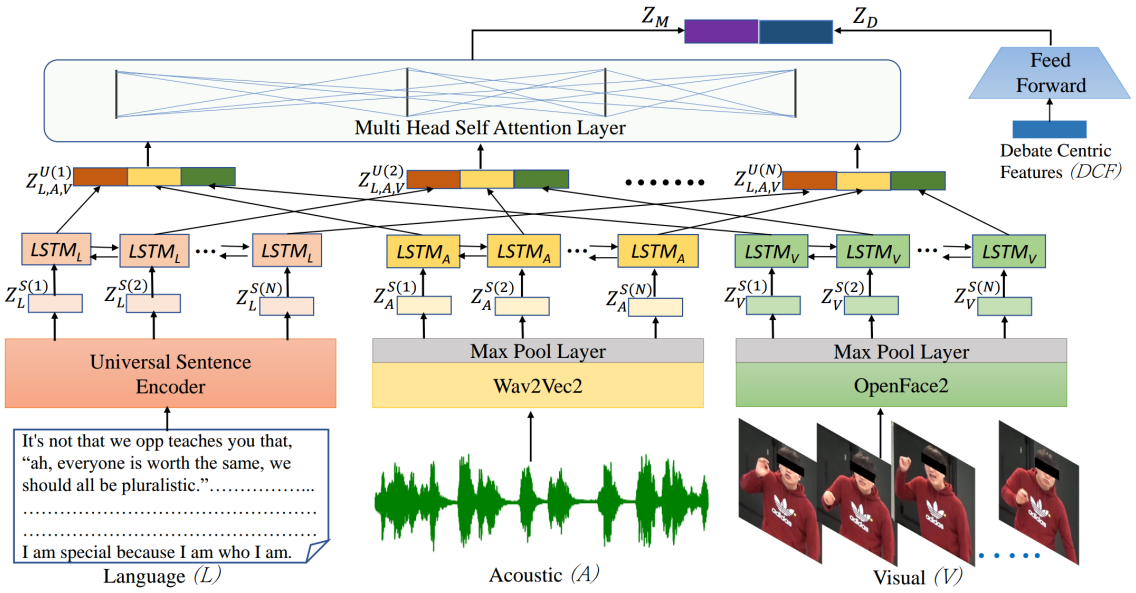

Figure 2: Overview of MARQ Model

et al., 2020). To measure the movement, we compute the euclidean distance of the body landmark points in consecutive frames. The mean values of these distances for all the landmark points are taken as an approximation of the body movement.

4 MARQ Neural Model

Each data point in the DBATES multimodal dataset can be represented as $X_i = \{L, A, V, DCF\}$ where L= language, A=acoustic, V= visual and DCF= debate-centric features. In addition to the debate-centric features (DCF), we consider the raw text (L), acoustic (A) and visual (V) information to model the summary of the video content. Given these features our task is to predict whether the arguments presented in the debate are of high quality or not. Each debate is around 6 minutes in length and the average text length is around 2000 words. To model this long multimodal sequence, we took the hierarchical approach in our MARQ model (Figure 2). First, we extract the sentence level embeddings from the pre-trained models. Then a system of LSTM encoders is used to learn the temporal relations in the unimodal sentence embeddings. Finally, the unimodal sentence embeddings go through a Multi Head Self Attention layer to create multimodal representation and enrich it with debate-centric features.

4.1 Sentence Level Representations

Language: Assume N the number of sentences in a debate video $L = [L_1, L_2, ..., L_N]$. Universal Sentence Encoder is used to extract embeddings of each sentence (Cer et al., 2018). The sentence level embeddings of language can be represented as $Z^S_L = UniversalSentenceEncoder(L)$; where $Z^S_L \in R^{N \times d^s_l}$ and d^s_l= dimension of the universal sentence encoder embedding. We also use Sentence-BERT (Reimers and Gurevych, 2019) to extract sentence embeddings and experiment with both variations.

Acoustic: The wav2vec2 (Baevski et al., 2020) is a pretrained transformer model of speech recognition that learns the representations of raw audio in self-supervised manner. It converts the speech input into discrete latent representations and learns the contextual representations via contrastive task. We use the base model that was trained on the 960 hours of Librispeech data (Panayotov et al., 2015). To extract the sentence level acoustic representations, the input audio file of the debate is split into N (#sentence) segments $A = [A_1, A_2, ..., A_N]$. The pretrained wav2vec2 model takes the raw audio segment of a sentence i and outputs contextual latent representations. These latent representations of sentence i go through a MaxPool layer. The max-pooling gives us computationally efficient method of extracting the most salient features across the time dimension and yields a fixed dimensional vector. This fixed dimensional vector is the sentence level acoustic embedding of sentence i. The sentence level acoustic representations can be represented as $Z^S_A = MaxPool(Wav2Vec2(A))$; where $Z^S_A \in R^{N \times d^s_a}$, d^s_a = the hidden dimension of

the wav2vec2 model.

Visual: OpenFace2 (Baltrusaitis et al., 2018) is used to extract facial Action Units (AU) features and rigid and non-rigid facial shape parameters. Facial action unit features are based on the Facial Action Coding System (FACS) (Ekman, 1997) which are widely used in human affect analysis. For each frame we extract these features using OpenFace2. To create sentence level embeddings, we take all the feature vectors from the frames of a sentence and apply max-pooling to get fixed dimensional (d_v^s) vector. The sentence level visual representations can be represented as $Z_V^S = MaxPool(OpenFace2(V))$; where $Z_V^S \in R^{N \times d_v^s}$, d_v^s = the dimension of the OpenFace2 features.

4.2 Unimodal Representation of Sentences

The sentence level representations of the language (Z_L^S) , acoustic (Z_A^S) and visual (Z_V^S) are extracted independently. To learn the temporal relations among the N sentences, three Bidirectional LSTM are used for each modality. The outputs of the three LSTM create the unimodal representations of the respective modalities. The unimodal representations of language can be denote as $Z_L^U = LSTM_L(Z_L^S)$; where $Z_L^U \in R^{N \times d_l^u}$ and d_l^u= is the hidden dimension of the $LSTM_L$. Similarly the unimodal representations of acoustic and visual are : Z_A^U; where $Z_A^U \in R^{N \times d_a^u}$ and Z_V^U; where $Z_V^U \in R^{N \times d_v^u}$.

4.3 Multimodal Representation Learning

A multihead self attention layer (Vaswani et al., 2017) is used to learn the inter modal interactions among language, acoustic and visual. The self-attention heads calculate the weighted summation of values (V); where the weights are computed from the scalar dot product of query (Q) and key (K) vector.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_h}})V \quad (1)$$

Multiple self-attention heads operating in parallel create Multi-Head Self Attention Layer - each potentially focusing on complementary aspects of the multimodal input. First, we concatenate the unimodal representations of the language, acoustic and visual. So, $Z_{L,A,V}^U = Z_L^U \oplus Z_A^U \oplus Z_V^U$; where $\oplus$ represents the concatenation and $Z_{L,A,V}^U \in R^{N \times (d_l^u + d_a^u + d_v^u)}$. Then it goes through a Multi-Head Self Attention Layer that learns the interaction across the modalities and output multimodal representations. The output of the Multi-Head Self Attention Layer goes through a Max Pool layer to create fixed dimensional vector of multimodal representation (Z_M). $Z_M = MaxPool(MultiHeadSelfAttention(Z_{L,A,V}^U))$; where $Z_M \in R^{d_l^u + d_a^u + d_v^u}$.

4.4 Debate-Centric Feature Representation

The debate-centric features (DCF) are extracted on the entire debate video. These features go through a fully connected neural network to create non linear projections. $Z_D = \mathcal{F}(DCF)$; where $\mathcal{F}$ is a fully connected neural network. Finally, the multimodal representation (Z_M) and and the debate-centric feature representation (Z_D) get concatenated. The resulted representation is passed through a fully connected neural network and softmax layer to compute the output probability. $p = softmax(\mathcal{F}(Z_M \oplus Z_D))$. This probability is used to predict if the given debate video got the high performance score or not.

5 Experiments

In this section, we discuss the baseline models and the hyperparameter settings that are used in the experiments.

5.1 Baseline Models

Logistic Regression: In addition to the debate-centric features (DCF), the average sentence level representations of the language(Z_L^S), acoustic (Z_A^S) and visual (Z_V^S) are used as feature for the logistic regression. We also train logistic regression with DCF features only to assess the importance of different debate-centric features.

MulT (Multimodal Transformer for Unaligned Multimodal Language Sequences): It has a set of cross modal transformer encoders that captures the bimodal interaction between the modalities. Then it summarizes all biomodal information to model the multimodal sequence (Tsai et al., 2019).

FMT (Factorized Multimodal Transformer for Multimodal Sequence Learning): It uses seven distinct self-attention heads to model the multimodal dynamics in a factorized manner, capturing all possible uni-modal, bi-modal, and tri-modal interactions, simultaneously (Zadeh et al., 2019).

Both of these neural models achieve state of the art performance in multimodal sentiment and emo-

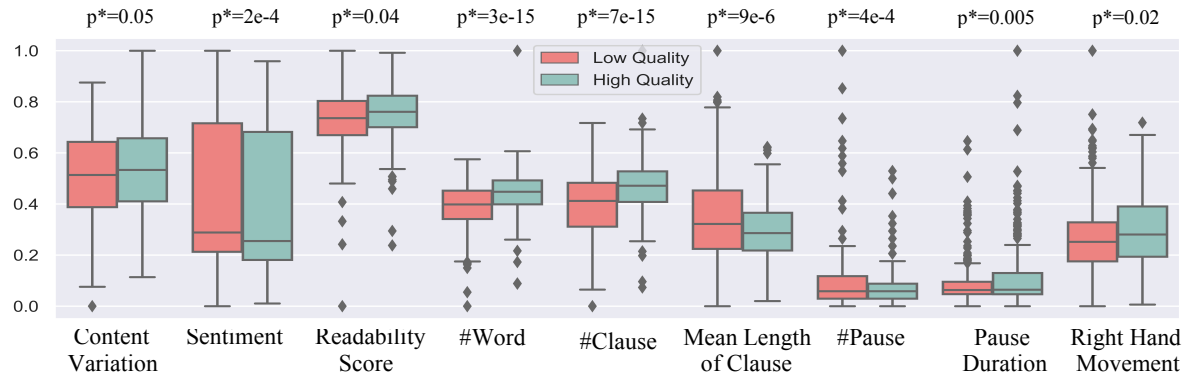

Figure 3: Examples of DCF feature distributions that achieve statistical significance ($p^* < 0.05$) in Student t-test analysis. Feature values are normalized in [0-1] range using min-max scaling.

tion prediction task. However, the complexity of transformer encoder increases exponentially with the length of the sequence. As our input sequence is around 1500 words, it is not feasible to train these models in an end-to-end manner on word level. We thus use sentence level embeddings (Z_L^S,Z_A^S and Z_V^S) that are extracted from the pretrained models as input to the MulT and FMT models.

5.2 Hyper Parameter Search

The dataset is divided into 75:15:15 split and the splits are speaker independent. Binary cross entropy is used as loss function. We experiment with {32,56,128,256,512} hidden dimensions for the unimodal LSTM's, [2-11] heads for the MultiHead Self Attention Layer, {0.05, 0.08, 0.1,0.2,0.15,0.3,0.25} dropout values and {1e-2,e-3,e-4} learning rates. The search space of the baseline models are given in the supplementary materials. All experiments are run on K80 gpus.

6 Results & Discussion

We analyze the statistical significance of the debate-centric features and then compare our MARQ model with established multimodal baselines.

6.1 Interpretability of the Debate-Centric Features

In this experiment, we assess whether the debate-centric features can capture meaningful patterns associated with argument quality. For each debate-centric feature, Student t-test analysis is performed to observe whether the feature plays a significant role in the differences between the distribution of high and low-quality arguments. We present features with statistical significance ($p < 0.05$) in **Figure** 3. The complete list of significant features is provided in the supplementary materials.

High-quality arguments show higher ($p = 0.05$) content variation compared to low-quality arguments. It indicates that the high-performing debaters speak with more diversity compared to the low-scoring debaters who show less content variation, possibly resulting in a monotonous delivery.

An interesting finding is that high performing debaters express more ($p = 2e^{-4}$) negative sentiments in their speech. This may suggest that in the context of a debate, the debaters often use negative words to expose the weakness of the opposition's stance. Using strong emotional expression (although negative) might increase the credibility of their stance on the debate topic.

The clarity of an argument makes it easy to persuade the audience. This is backed up by our finding that high-quality arguments have a higher ($p = 0.04$) readability score compared to low-quality arguments. Good arguments have simple sentence structures that are easy to understand. From the syntactic complexity features, we observe that debaters with higher scores use more words and clauses in their speech ($p = e^{-15}$). It is possible that the low-scoring debaters struggle to find enough content for their arguments. The good debaters also use fewer ($p < 0.05$; not listed in Fig 3) complex nominal and coordinate phrases per clause and avoid complex sentence structures (short clauses) to make their arguments clear.

We find that debaters with low scores take more ($p = 4e^{-4}$) pauses than debaters with high scores. However, the average duration of the pause is longer ($p = 5e^{-3}$) among the debaters with high scores. A possible explanation is that the low scorer debaters hesitate during their argument delivery that resulted in more unintentional short pauses, making it difficult for the audience to pay full attention. Perhaps, the good debaters have more control

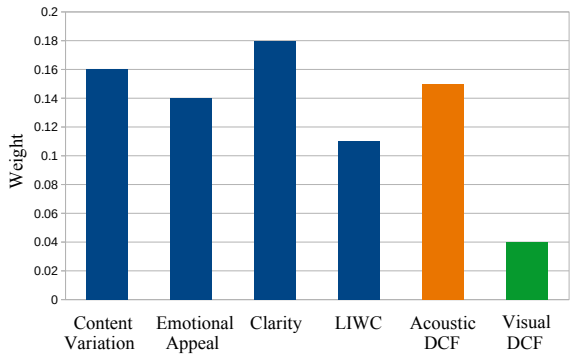

Figure 4: Relative importance of different sub-categories of debate-centric features obtained from the Logistic Regression Model

in their speech and they plan their pauses, allowing audience to follow along. The hand and gesture movement of the debaters is also correlated with the scoring of their argument ($p = 0.015$). We find that debaters who move their right hand more often usually scored higher. However, we could not find any significance of the left-hand movements. It is possible that the number of left-handed debaters in this dataset was not large enough to provide any statistical significance.

We also run logistic regression using debate-centric features only to distinguish the high-quality arguments from the low-quality ones. It achieves an F1-score of 75.23%, which demonstrates the discriminative power of these features. The relative importance of the debate-centric features is analyzed from the associated weights of the logistic regression. The cumulative weights of each sub-category are normalized by the number of features from each sub-category (Figure 4). We observe that **Clarity** is the most important feature for distinguishing the argument quality while **Content Variation** and **Acoustic-DCF** features also play a key role in the classification. We find that **Visual-DCF** has the least importance. This is possible because the judges were specifically instructed not to be biased by visual appearances.

These interpretable features not only help the model to achieve good performance, but also reveal debating strategies to deliver more convincing arguments. Models trained on these features will be helpful to design feedback system for training debaters to improve their argumentation skill.

6.2 Argument Quality Prediction

The results of the multimodal argument quality prediction are presented in Table 2. A Logistic Regression model that uses all the modalities and DCF features achieves 75.53% accuracy and 76.77% F1 score respectively in predicting high vs low-quality arguments.

Models	Accuracy	F1 Score
Logistic Regression	75.53	76.77
MulT	76.60	76.57
FMT	69.15	67.70
MARQ (Sentence Bert)	79.79	79.78
MARQ	**81.91**	**81.88**

Table 2: Performances of the competitive models on multimodal argument quality prediction task. Binary accuracy and F1-score are reported here as performance metrics.

Models	Accuracy	F1 Score
MARQ	**81.91**	**81.88**
- DCF	78.72	78.43
- acoustic	80.85	80.83
- visual	80.85	80.85

Table 3: Role of modalities in our MARQ model. Here '-' denotes removal of the corresponding feature set.

Multimodal Transformer model (MulT) (Tsai et al., 2019) achieves 76.60% accuracy and 76.57% F1 score on this task. It has six cross-modal transformer-based encoders to capture the cross-modal interactions and three transformer encoders to fuse the cross-modal information. It overfits quickly due to a high number of parameters. A similar trend is also observed in the FMT model (Zadeh et al., 2019) since it also has a large number of parameters. Moreover, the FMT model performs worse than the logistic regression baseline indicating the limitation of this model in a low resource scenario.

Our MARQ model outperforms all of these established baselines by achieving 81.91% accuracy (22.7% error rate reduction compared to the MulT model) and 81.88% F1 score (22.7% error rate reduction compared to MulT). The Sentence-BERT variation of MARQ does not achieve similar performance, possibly because it was not pre-trained on a related task.

Finally, we study the role of different modalities by re-training the MARQ model after removing all features of a modality (one modality at a time). The performance is reported in Table 3. Removing debate-centric features has the worst impact, increasing the error rate by 17.63%. Visual and acoustic modalities have a similar impact on the multimodal argument quality prediction. Though the MulT and FMT models do not use DCF fea-

tures, they perform worse than the MARQ model not using DCF features. This indicates the importance of MARQ type architecture in modeling a low-resource multimodal dataset of long sequences.

7 Related Work

The automatic assessment of argument quality (Toledo et al., 2019; Gienapp et al., 2020) has been receiving growing interest in the NLP community. Identifying argument quality has applications in diverse domains, including but not limited to argument search (Wachsmuth et al., 2017b,c), finding counter arguments (Wachsmuth et al., 2018), automated decision making (Bench-Capon et al., 2009), writing support (Stab and Gurevych, 2014) and essay evaluation (Nguyen and Litman, 2018). Wachsmuth et al. (2017a) proposed a taxonomy of dimensions for quantifying argument quality, where they summarized several high level dimensions behind the structure of good arguments such as clarity, coherence, effectiveness, emotional appeal, etc. However, the subjective nature of these dimensions makes the task of automatic argument quality scoring difficult.

Earlier research on automatic argument quality assessment focused on comparative pairwise approach, where the task is to identify higher quality argument from a given pair of arguments (Habernal and Gurevych, 2016b; Simpson and Gurevych, 2018; Potash et al., 2019; Gleize et al., 2019). Recently, Toledo et al. (2019) introduced straightforward point-wise argument quality metric that scales with the data size linearly. They introduced IBM-RANK (6.3K text arguments) that was crowd sourced and then annotated with an individual quality score. Following similar approach, Gretz et al. (2020) proposed IBM-RANK-30k – the largest dataset of argument quality score prediction in free text. Both of them utilized BERT (Devlin et al., 2018) based fine tuning for this task.

The previous research and datasets (Table 1) are mostly limited to short text sequences (18-48 words). Also, most of the prior work only consider a single modality (text). However, real life arguments are multimodal. Non-verbal cues like facial expression, body language, prosodic strategies often amplify or dampen the quality of a given argument. Analysis based on the unimodal signal is not fully inclsuive of real-world characteristics and could lead to misleading findings (Braga and Marques, 2004; Straßmann et al., 2016; Hasan et al., 2019c). That's why there exists vast amount prior research that utilize multimodal data to understand human communication behavior properly (Rahman et al., 2020; Hasan et al., 2021; Zadeh et al., 2018a; Tsai et al., 2019; Samrose et al., 2019; Sen et al., 2018; Hasan et al., 2019b; Zadeh et al., 2018b; Hasan et al., 2019a). Petukhova et al. (2017) discuss the design and evaluation of a Virtual Debate Coach (VDC) for training young politicians to improve their debate skills. They used logistic regression to identify multimodal features correlated with debate performance. Their DTC dataset comprised of 400 debate videos collected from professional debaters. Another similar work (Hirata et al., 2019) also uses logistic regression of multimodal features to assess argument quality and thereby generate automated feedback. However, none of the above studies released their dataset for further research.

Recently, Sen et al. (2021) publicly released DBATES – a dataset of debate videos ($N = 716$) collected from the 2019 North American Universities Debate Championships. The authors performed logistic regression to show that beside text, other nonverbal features have correlation with the performance of a debate. The DBATES dataset also presents a global challenge applicable to any multimodal assessment task – representing multimodal signals in a long video, which was not addressed by the authors. In this study, we use this dataset and make two major contributions – 1) first to study multimodal argument quality assessment beyond logistic regression; 2) address a technical challenge of multimodal representation for long videos (6 minutes on average).

8 Conclusion

In this paper, we presented a comprehensive study on multimodal argument quality assessment. The debate-centric features reveal interpretable patterns associated with the quality of argument and help improve the prediction performance. These features can easily be adapted to a working system with transparent, objective, repeatable feedback on assessing the quality of a speech and its arguments, and thus lead to equitable access to a training system for anyone wanting to become a good debater. We also proposed a hierarchical neural model (MARQ) to assess the quality of argument in a long video and showed the importance of having nonverbal cues through further ablation studies.

Although our work is limited to the only publicly available video dataset of debate, we hope it will inspire others to study the task of argument quality assessment in multimodal context, and develop new datasets and algorithms.

The code and data described in this paper are publicly available at https://github.com/matalvepu/MARQ

Acknowledgments

This research was supported in part by grant W911NF-19-1-0029 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO), and the National Science Foundation NRT-DESE 1449828.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Trevor Bench-Capon, Katie Atkinson, and Peter McBurney. 2009. Altruism and agents: an argumentation based approach to designing agent decision mechanisms. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1073–1080. Citeseer.

Daniela Braga and Maria Aldina Marques. 2004. The pragmatics of prosodic features in the political debate. In *Speech Prosody 2004, International Conference*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Justin Eckstein and Michael Bartanen. 2015. British parliamentary debate and the twenty-first-century student. *Communication Studies*, 66(4):458–473.

Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Rudolf Flesch and Alan J Gould. 1949. *The art of readable writing*, volume 8. Harper New York.

Michelle Fung, Yina Jin, RuJie Zhao, and Mohammed (Ehsan) Hoque. 2015. Roc speak: Semi-automated personalized feedback on nonverbal behavior from recorded videos. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 1167–1178, New York, NY, USA. Association for Computing Machinery.

Ahmed Gheith, Ramakrishnan Rajamony, P Bohrer, Kanak Agarwal, Michael Kistler, BL White Eagle, CA Hambridge, John B Carter, and T Kaplinger. 2016. Ibm bluemix mobile cloud services. *IBM Journal of Research and Development*, 60(2-3):7–1.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Ivan Habernal and Iryna Gurevych. 2016a. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Kamrul Hasan, Wasifur Rahman, Luke Gerstner, Taylan Sen, Sangwu Lee, Kurtis Glenn Haut, and Mohammed Hoque. 2019a. Facial expression based imagination index and a transfer learning approach to detect deception. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 634–640. IEEE.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12972–12980.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019b. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Md Kamrul Hasan, Taylan Sen, Yiming Yang, Raiyan Abdul Baten, Kurtis Glenn Haut, and Mohammed Ehsan Hoque. 2019c. Liwc into the eyes: Using facial features to contextualize linguistic analysis in multimodal communication. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.

Hayato Hirata, Shogo Okada, and Katsumi Nitta. 2019. Analysis of argumentation skills for argumentation training support. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 319–334. Springer.

Kellogg W Hunt. 1965. A synopsis of clause-to-sentence length factors. *The English Journal*, 54(4):300–309.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Huy Nguyen and Diane Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Volha Petukhova, Tobias Mayer, Andrei Malchanau, and Harry Bunt. 2017. Virtual debate coach design: assessing multimodal argumentation performance. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 41–50.

Peter Potash, Adam Ferguson, and Timothy J Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*, pages 146–155.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

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

Andrew Rosenberg and Julia Hirschberg. 2009. Charisma perception from text and speech. *Speech Communication*, 51(7):640–655.

Samiha Samrose, Wenyi Chu, Carolina He, Yuebai Gao, Syeda Sarah Shahrin, Zhen Bai, and Mohammed Ehsan Hoque. 2019. Visual cues for disrespectful conversation analysis. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 580–586. IEEE.

Taylan Sen, Md Kamrul Hasan, Zach Teicher, and Mohammed Ehsan Hoque. 2018. Automated dyadic data recorder (addr) framework and analysis of facial cues in deceptive communication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22.

Taylan K Sen, Gazi Naven, Luke Gerstner, Daryl Bagley, Raiyan Abdul Baten, Wasifur Rahman, Kamrul Hasan, Kurtis G Haut, Abdullah Mamun, Samiha Samrose, et al. 2021. Dbates: Database of audio features, text, and visual expressions in competitive debate speeches. *arXiv preprint arXiv:2103.14189*.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Carolin Straßmann, Astrid Rosenthal von der Pütten, Ramin Yaghoubzadeh, Raffael Kaminski, and Nicole Krämer. 2016. The effect of an intelligent virtual agent's nonverbal behavior with regard to dominance and cooperativity. In *International Conference on Intelligent Virtual Agents*, pages 15–28. Springer.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. "pagerank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2236–2246.