Simple Attention-Based Mechanism for Multimodal Emotion Classification

Mazen Elabd and Sardar Jaf
School of Computer Science
University of Sunderland
Sunderland, United Kingdom
bi24bh@student.sunderland.ac.uk
sardar.jaf@sunderland.ac.uk

***Abstract*—Recent advances in attention-based machine learning models have significantly enhanced performance in various classification tasks, including emotion recognition. In this work, we introduce two novel multimodal architectures that leverage attention-based techniques for emotion classification. We introduce a state-of-the-art attention-based multimodal architecture and a baseline architecture. Our state-of-the-art architecture utilises attention-based unimodal models to extract contextualised embeddings from each modality and an attention-based fusion technique. Performance metrics and rigorous error analysis indicate that unimodal systems trained solely on text or speech data underperform compared to our multimodal system, which integrates both modalities. Furthermore, our system significantly outperforms existing state-of-the-art multimodal systems using the same modalities by nearly 4%.**

***Index Terms*—emotion classification, emotion recognition, multimodal emotion classification, multimodal fusion, multimodal classification.**

I. Introduction

Recognising human emotions automatically from text is a challenging task. Textual data lacks emotional cues, such as speech tone, pitch, vocal expression, and facial expression, which are helpful in accurately determining a person's emotion. Therefore, approaches to automated emotion recognition that rely only on text are inherently limited. Recent attempts at emotion recognition have focused on using additional types of information such as audio, image, and video along with text to enrich the information needed for accurately classifying human emotions.

With advances in machine learning applications for various tasks (such as image processing, computer vision, and natural language processing), it is now possible to design homogeneous multimodal classification systems that incorporate state-of-the-art machine learning unimodal models trained on labelled datasets containing multiple types of data modalities such as text, images, audio, and video. Training a multimodal emotion classifier can involve using one or more machine learning algorithms that are best suited for learning from different types of data. For example, we could train algorithm x on text data, y on audio data, and z on visual data, and then integrate their outputs to perform the classification task. In this paper, we propose a novel deep-learning architecture that uses text and audio information for the task of emotion classification. Our contributions are as follows:

- We fine-tune two unimodal transformer-based models for text and audio emotion recognition.
- We propose a novel baseline deep-learning multimodal architecture to extract and utilise important features from different types of data (text and audio) to classify the input into one of several emotion classes.
- We propose a new state-of-the-art multimodal emotion classifier.

The rest of the paper is organised as follows: Section II outlines key related work relevant to our proposed system. We describe our methodology in Section III and in Section IV we present the performance of the proposed system and compare it against several published works. In Section V we analyse the system's performance through several confusion matrices. We conclude the paper in Section VI.

II. Related work

The advances in deep learning methods for audio and text processing have motivated researchers to develop various approaches to emotion classification. Generally, these approaches involve training deep learning models on audio and text data, and incorporating a fusion mechanism to combine both modalities. Early feature extraction methods from text, such as word2vec, focused on learning information from text based on word features. Subsequent advancements in feature extraction from text include bidirectional long short-term memory networks (Bi-LSTM), gated recurrent units (GRU), and transformers. Additionally, advanced language models that capture multilingual and contextual information around words have been developed such as Bidirectional Encoder Representations from Transformers (BERT) [1], Robustly optimized BERT (RoBERTa) [2], and GPT [3].

Methods to combine multiple data modalities (e.g., text, audio, and images) have proven effective in emotion classification [4]. Multimodal transformers, which enable the concatenation of feature representations from various data types have replaced previous methods [5] [6]. Further advances include multi-view sequential learning models [7] and dynamic fusion graphs [8]

Dutta et al [9] proposed a hierarchical cross-attention model approach using recurrent and co-attention neural network models trained on text and audio data. Their first stage involved training an utterance-level embedding extractor from the input data (text and audio), which trains the model to classify individual utterances without accounting for the inter-utterance conversational context. The second stage involves feeding the utterance-level embeddings from the first stage into a bidirectional gated recurrent unit (Bi-GRU) to incorporate inter-utterance context into the model. This stage enriches the model with conversational context information. The final stage involves fusing the embeddings from different modalities using self-attention and cross-attention mechanisms. These attention mechanisms allow for capturing relationships and dependencies within input sequences (sentences or speech utterances).

[10] proposed a hierarchical transformer-based model (HiTrans) consisting of a transformer-based content model and a speaker-sensitive model for emotion classification. Their method uses two hierarchical transformers: a BERT model is used as the low-level transformer for generating local utterance representations, and a high-level transformer takes the output of the low-level transformer as input to make the model sensitive to the global context of the conversation. They integrate a "pairwise utterance speaker verification" (PUSV) method to detect whether two utterances belong to the same speaker.

[11] proposed the Dialogue Graph Convolutional Network (DialogueGCN) model utilising a graph neural network approach to emotion classification. It leverages self and inter-speaker dependencies among interlocutors to model conversational context. One of the main advantages of this method is its ability to address context propagation issues present in current RNN-based methods.

[12] proposed a learning-based system for emotion classification using multiple input modalities that combine information from text, facial cues, and speech. Their system seems to pay more attention to reliable cues while suppressing less helpful cues on a per-sample basis by using Canonical Correlation Analysis, which differentiates between effective and ineffective modalities. The major strength of their proposed system is its robustness to sensor noise in any of the individual modalities.

III. Proposed Method

We propose a supervised machine-learning approach for emotion classification. We utilise a public dataset containing various data types (text, speech, and video) to train and evaluate a multimodal emotion classifier. The benefit of using different types of data is to extract different levels of details and cues for identifying emotions. Each data type provides unique information for training machine learning algorithms. Due to limited computational resources, we only use text and speech data in this study. For text, we utilise BERT contextualised embeddings [1] and for speech, we utilise Audio Spectogram Transformer (AST) contextualised embeddings [13].

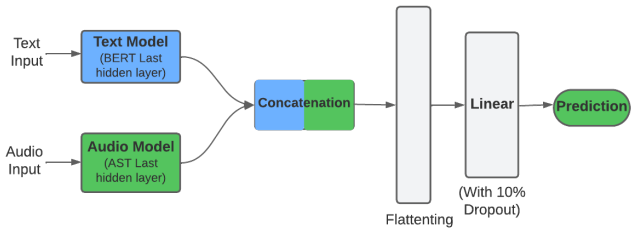


Fig. 1: Baseline multimodal classification architecture.

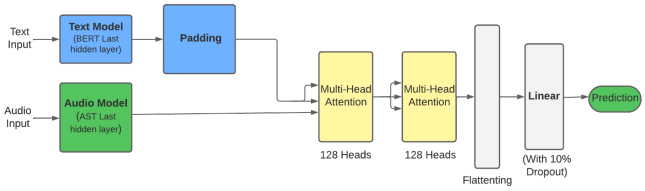


Fig. 2: Multi-head attention-based multimodal architecture.

A. System Design

Our system design is based on fine-tuning attention-based models as well as the reliance on attention-based feature fusion. We believe that two main problems should be addressed when building any multimodal architecture: i) the data representation for each modality in the system (i.e., data representation), and ii) the data concatenation from different modalities (i.e., data fusion).

Our proposed multimodal system for emotion classification relies on the fusion of extracted features from the data through two unimodal models: a BERT-based model for text data, and an AST-based model for speech data. The last hidden layer from both BERT and AST models is known to have the most contextualised representation of the input sequence. Thus, we specifically use the output of the last hidden layer of fine-tuned versions of BERT and AST to represent the text and audio modalities, respectively. In our attention-based architecture, we adjust the size of the output of the last hidden layer from the text model through a padding layer to match the size of the output of the last hidden layer for the audio model, as in Fig. 2. Similar to late-fusion-based multimodal systems, relying on fine-tuning pre-trained transformer-based models gives the capability of utilising powerful and highly efficient models to represent the multimodal data with a relatively small dataset size and limited hardware resources in the training process.

To address the feature fusion problem, we rely on a multi-head cross-attention layer as a concatenation technique. Using a method such as averaging or voting after the unimodal classifier layers would prevent the multimodal system from benefiting from potential interactions between different modalities.

Instead of merging both modalities at a classifier level, as done in late feature fusion techniques, we decided to

merge them at a data representation layer, which is similar to early fusion techniques, to be able to capture the interactions between audio and text. Our proposed system leverages the best from both early and late fusion techniques. We use powerful pre-trained unimodal models while merging the data before the classifier layer using a simple concatenation layer (as shown in Fig. 1 for the baseline system) and cross-attention fusion technique, as presented in Fig. 2. We believe that cross-attention fusion is capable of capturing the interactions between different modalities. It may also be able to make sense of confusing and contradicting data.

To perform the cross-attention fusion, we feed the output from the last hidden layer of the audio model as the key tensor to the multi-head attention layer while feeding the padded output from the last hidden layer of the text model as the query and the value tensors of the multi-head attention layer, as shown in Fig. 2. The multi-head cross-attention layer is followed by a multi-head self-attention layer which we added in the pursuit of enriching the combined audio and text representation. The combined feature vector is fed as the query, key, and value for the multi-head self-attention layer, as in Fig. 2. After combining the features from the different data modalities, the combined feature vector undergoes flattening along the first dimension. Then, the combined feature vector is passed to a linear layer with a 10% dropout, which is followed by the classification layer, as in Fig. 1 and 2.

1) Selecting the unimodal models: Since the presented architectures in this work rely on unimodal systems to create feature representations for each modality, the selection of the unimodal system is critical to optimise our system's performance. We believe that the various unimodal models utilised within a multimodal classification architecture should rely on similar feature representation algorithms to reach a homogeneous feature vector when combining them in the system. We choose attention-based unimodal approaches to build the feature vector for each modality since they are proven to achieve state-of-the-art results in many benchmarks and most importantly to maintain homogeneous feature representations between modalities. Attention-based models are popular in various unimodal classification domains, such as text, speech, video, and image, which enable our multimodal systems to easily scale to include another modality. The selected unimodal systems are mainly based on transformer-encoder and only rely on attention to keep the models as homogeneous as possible. In addition, the selected systems are simple and fast to fine-tune.

The last hidden layer from each of the transformers (BERT and AST) contains the most comprehensive and contextualised representation of the input sequence; therefore, it is logical to use it as the feature vector for each modality.

2) Baseline system: For our baseline model, as depicted in Fig. 1, the information derived from the last hidden layer of BERT and AST is fed into a simple concatenation layer as the fusion layer. This was followed by a flattening layer, then a linear layer with dropout (10% dropout rate) as the classification layer.

TABLE I: The learning rate and the batch size used in each experiment

Model	Learning Rate	Batch Size
BERT	5e-6	2
AST	5e-6	2
Baseline	5e-8	2
Attention-based	5e-8	2

3) Attention-based system: For this system, we rely on feeding the feature vector from BERT and AST models to a multi-head cross-attention layer, as in Fig. 2, which acts as a fusion layer that is capable of capturing the interactions between modalities and developing a comprehensive understanding of the multimodal data. The multi-head cross-attention layer is followed by a multi-head self-attention layer to enrich the combined feature vector before classification. Next, a flattening layer and a linear layer with a 10% dropout rate serve as the classification layer.

B. dataset

We use Multimodal EmotionLines Dataset (MELD) [14] for the evaluation of the proposed systems. The dataset contains more than 13000 utterances derived from multi-party conversations. it is divided into three parts: training, validation, and testing. The training set consists of 9,988 utterances, the validation set 1,108 utterances, and the test set 2,610 utterances. We use these partitions unchanged for training, validation, and testing.

C. Hyperparameters

The MELD dataset suffers from imbalance class distribution which impacted our decision to rely on AdamW as an optimizer, the weighted cross-entropy as the loss function, and the weighted F1 score as our main evaluation metric. We integrated early stopping in all of our experiments to stop the training once the weighted F1 score started to decrease.

Some hyperparameters were only shared across the multimodal experiments such as using a 10% dropout rate in the last linear layer. Also, both multi-head attention layers used in this work were configured to consist of 128 attention heads. We performed limited experiments to reduce the number of attention heads, although the weighted F1 score was negatively impacted.

Despite the fact that we managed to achieve state-of-the-art performance, the selected hyperparameters still might not be perfect since we were only aiming to get an acceptable learning curve. Also, some hyperparameters, such as learning rate and batch size in table Table I, were configured due to hardware limitations as we only depended on a 12GB NVIDIA GeForce GPU to carry out the experiments in this work.

1) evaluation metric: Some of the most robust and widely used performance measures for classification tasks are: (i) recall, (ii) precision, and (iii) F1 score. The recall metric measures the proportion of actual positives that the model is able to classify. Precision, on the other hand, measures the proportion of predicted positives that are correctly identified.

TABLE II: Performance of text and audio unimodal models by emotion class.

	text			speech		
Categories	precision	recall	F1 score	precision	recall	F1 score
Anger	0.497	0.487	0.492	0.387	0.249	0.303
Disgust	0	0	0	0	0	0
Fear	0.5	0.02	0.038	0	0	0
Joy	0.602	0.627	0.615	0.228	0.192	0.208
Neutral	0.741	0.847	0.790	0.577	0.757	0.655
Sadness	0.376	0.337	0.355	0.175	0.154	0.164
Surprise	0.654	0.537	0.590	0.314	0.231	0.266
Macro Avg.	**0.481**	**0.408**	**0.412**	**0.240**	**0.226**	**0.228**
Weighted Avg.	**0.623**	**0.654**	**0.633**	**0.411**	**0.464**	**0.429**

Individually, these metrics do not offer a complete view of model performance. The F1 score computes the harmonic mean of the recall and precision offering a robust evaluation. Since the data are imbalanced, we use a weighted F1 score to account for each class's contribution based on its support, which considers the number of actual occurrences of the class in the dataset. Moreover, we compute and present the macro average and weighted average performances of the models to show the overall system performance. We present the macro average to assess precision, recall, and F1 score across individual classes, treating all classes equally. We also use the weighted average performance to account for the importance/weight of each category when calculating the overall system performance.

IV. Result and Discussion

The proposed deep learning systems consist of appropriate deep learning algorithms for each data type. They are trained and tested on two types of data: text and speech data.

We present several performance aspects of the proposed systems for recognising different types of emotions. We use several standard evaluation metrics for each category: precision, recall and F1 score, macro average and weighted average F1 score. These performance measures allow us to identify the model's performance at the category level. In the following subsections, we outline the performance of the systems when trained and tested for emotion classification.

A. Model evaluation result at category level

1) Text unimodal performance: Table II shows the performance of the text-based unimodal system in classifying emotions when trained and tested on text data. The second, third, and fourth columns show the precision, recall and F1 score for each type of emotion. The text-based unimodal model failed to classify the "disgust" emotion. Although it scores a recall of 50% when classifying the "fear" category, it has a very poor precision of 0.02 (2%), resulting in a very poor F1 score of 0.038. The text-based unimodal model classifies the "neutral" category more accurately than other categories, achieving an F1 score of 0.790. This is followed by the "joy" category (with an F1 score of 0.615) and the "surprise" category (with an F1 score of 0.590). The performance in classifying "anger" and "sadness" categories are F1 scores of 0.492 and 0.355, respectively.

TABLE III: Performance of the baseline multimodal model by emotion class.

Categories	precision	recall	F1 score
Anger	0.577	0.435	0.496
Disgust	0.5	0.012	0.029
Fear	0.154	0.08	0.105
Joy	0.568	0.657	0.609
Neutral	0.793	0.766	0.779
Sadness	0.333	0.375	0.353
Surprise	0.483	0.705	0.573
Macro Avg.	**0.487**	**0.433**	**0.421**
Weighted Avg.	**0.640**	**0.635**	**0.627**

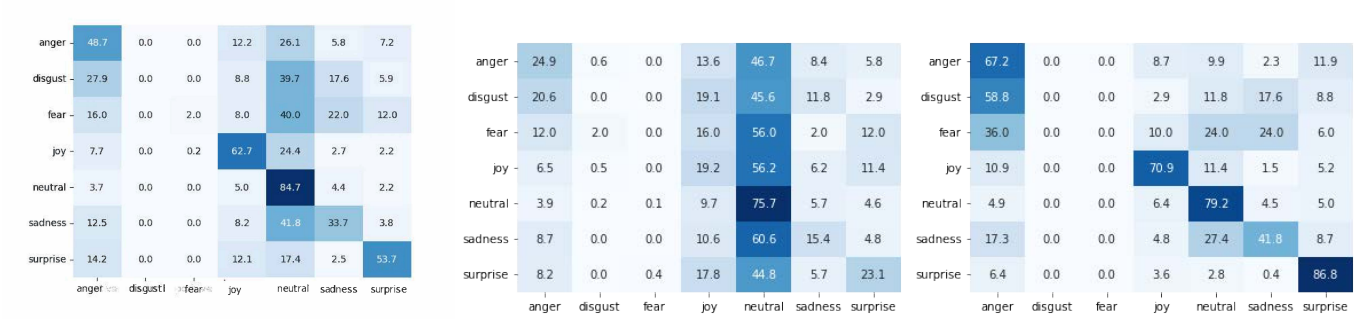
2) Speech unimodal performance: The speech-based unimodal system performance is relatively similar to that of the text-based unimodal system. It performs better in classifying the "neutral" emotion than any other type of emotion and performs very poorly in classifying "disgust" and "fear" emotions. As presented in Table II, the speech-based unimodal system fails to classify "disgust" and "fear" categories while achieving an F1 score of over 0.655 for the "neutral" category. Similar to the text-based unimodal system, following the "neutral" category, the speech unimodal system achieves an F1 score of 0.303, 0.266, and 0.208 for the "anger", "surprise", and "joy" categories, respectively.

3) Baseline multimodal performance: We evaluated the proposed multimodal baseline system to assess its performance at classifying different types of emotions. Table III presents the performance of the baseline system. The system produced the highest F1 score in classifying the "neutral" emotion (0.779), followed by "joy" (0.609). The system produced the lowest F1 scores in classifying the ' disgust" and 'fear" categories with 0.029 and 0.105, respectively. The overall F1 score across all the emotion categories is 0.627. The baseline system has a reasonable overall weighted average of precision and recall across all emotion categories, producing scores of 0.640 and 0.635, respectively.

4) Attention-based multimodal performance: Table IV presents the performance of the attention-based system. The system produced the highest weighted F1 score, recall, and precision compared to other models in this work, with scores of 0.693, 0.706, and 0.689, respectively. It also produced the highest macro average F1 score, with 0.467. Although it was unable to classify any instances from the least represented classes "fear" and "disgust", the system produced the highest F1 score in classifying the "neutral" emotion (0.824), followed by "surprise" (0.721), and then "joy" (0.692).

B. System performance comparison with previous works

Table V presents the weighted F1 score performance of our proposed systems (baseline and attention-based) and several previously published works on unimodal and multimodal emotion classification. Our proposed unimodal systems lagged behind some previously published systems, while our proposed multimodal systems have performed better than most of the other systems.



(a) Text unimodal model (b) Speech unimodal model (c) Attention-based multimodal model

Fig. 3: Confusion matrix graphs in the emotion classification task.

TABLE IV: Performance of the attention-based multimodal model by emotion class.

Categories	precision	recall	F1 score
Anger	0.516	0.672	0.584
Disgust	0	0	0
Fear	0	0	0
Joy	0.675	0.709	0.692
Neutral	0.858	0.792	0.824
Sadness	0.478	0.418	0.446
Surprise	0.616	0.868	0.721
Macro Avg.	**0.449**	**0.494**	**0.467**
Weighted Avg.	**0.689**	**0.706**	**0.693**

TABLE V: Comparative result. Weighted F1 score (in %)

Models	**Speech**	**Text**	**Speech+Text**
Majumder et al [15]	41.80	57.00	60.30
Ho et al [16]	45.30	59.00	60.50
Lian et al [17]	38.20	58.30	60.50
Baijun et al [18]	32.10	61.20	64.00
Dutta et al [9]	50.10	65.60	65.80
Proposed baseline model	42.90	63.3	62.7
Proposed attention-based model	42.90	63.3	**69.7**

Our proposed attention-based system significantly outperformed state-of-the-art systems by achieving a 69.7% F1 score, nearly 5% higher than the reported state-of-the-art system proposed by [9]. This is despite the low performance of our unimodal systems compared to the unimodal systems proposed by [9]. This indicates the strength of our novel method of combining the last hidden layers of attention-based text and speech unimodal models using a multi-head cross-attention layer followed by a multi-head self-attention layer for fusing text and audio modalities.

C. Multimodal performance discussion

Our proposed speech-based unimodal emotion classifier underperforms compared to those classifiers proposed by Dutta et al [9] and Ho et al [16]. Our text-based unimodal model is only outperformed by the model proposed by Dutta et al [9]. However, our multimodal classifier outperforms the state-of-the-art. We anticipate that the superior performance of our multimodal system results from our homogeneous attention-based model which includes an attention-based fusion method that merges the learned information from attention-based unimodal models, which are fine-tuned versions of BERT and AST. In comparison to several other published works, our speech-based unimodal performance is on par with many other systems and better than others in some cases. Our text-based unimodal model performs better than almost all other published works, as shown in Table V. Furthermore, our multimodal attention-based system outperforms all other published works, making it the state-of-the-art multimodal emotion classifier.

V. Error analyses: confusion matrix

Machine learning applications for supervised classification tasks require a set of labelled data that represents the classes. Labelled datasets often contain errors (annotation errors), where some data samples are labelled with incorrect emotions. Annotation errors impact machine learning algorithms' performance because the algorithms learn from the errors present in the annotated dataset.

In addition to annotation errors, other causes for misclassification include class imbalance, an issue that is highly present in the MELD dataset, which might lead to a biased model performance favouring the most represented class, "neutral" in our case. Furthermore, we believe the MELD dataset might have insufficient training data for the least represented classes, which could lead to the absence of core class characteristics. Our model might be able to overcome the class imbalance issue through our hyperparameter selection, such as our use of AdamW as an optimizer and a weighted cross-entropy as the loss function. However, it is extremely difficult to overcome insufficient training data solely by tweaking the hyperparameters without enriching the data quality.

Confusion matrix graphs are helpful for highlighting classification errors by identifying misclassified classes and the classes they are confused with. We present several confusion matrices to outline the classification errors of our proposed unimodal and multimodal systems in classifying different types of emotions.

A. Emotion misclassification errors

Fig. 3 presents the misclassifications of our attention-based multimodal system when evaluated on the MELD dataset.

As presented in Fig. 3(a), the "anger" class is misclassified as "neutral" 26.1%, "disgust" is misclassified as "neutral" 39.7% of the time, "fear" is misclassified as "neutral" 40% of the time, and topped by "sadness" with 41.8%. Contrary to the "neutral" class, which is the most represented class in the dataset, the text-based unimodal system does not confuse any class (except for "joy" confused with "fear", though negligibly) with the "disgust" and "fear" classes, which are the least represented classes.

The speech-based unimodal system seems to misclassify most classes as "neutral", as shown in Fig. 3(b). The misclassification error rates between "neutral" and all other classes range from 44.8%, where "surprise" is misclassified as "neutral", to 60.0%, where "sadness" is misclassified as "neutral". The "neutral" class appears to have the least misclassification. It has 75.7% accuracy., with its highest misclassified rate of 9.7% as "joy" followed by "sadness" (5.7%) and "surprised" (4.6%).

The attention-based multimodal system, as presented in Fig. 3(c), confused the "disgust" class with "anger" (58.8% error rate), "neutral" (11.8% error rate), "sadness" (17.6% error rate), and "surprised 8.8% error rate). The attention-based multimodal has improved the accuracy of classifying the "anger" emotion by reducing the error rate of misclassification with "neutral" to 9.9% compared to the error rate recorded in the unimodal systems. However, the "anger" emotion is misclassified as "joy" 8.7%, as "sadness" 2.3%, and as "surprise" 11.1%. The "sadness" class has been predominantly confused by the attention-based multimodal system with "neutral" with an error rate of 27.4%, followed by 17.3% misclassification as "anger", 4.8% as "joy", and 8.7% as "surprise". Other classes ("joy" and "surprise") are mostly confused with "neutral" and the " neutral" class is largely confused with "joy" (6.4% error rate). It appears our proposed attention-based system performed best in classifying "surprise" emotion with 86.8% accuracy.

VI. Conclusion and Future Work

Recent advances in automated data classification of emotion involve the application of machine learning algorithms to automate the process of classifying certain types of emotions (e.g., anger, happiness, surprise, etc.). The focus of this study was to design and evaluate novel deep-learning architectures that learn from different types of data (text and speech) to classify several types of emotions. We proposed two unimodal models, a fine-tuned BERT and AST, and two multimodal models, a baseline and an attention-based model. Our novel attention-based multimodal system fuses the contextualised embeddings from attention-based unimodal models, BERT and AST, using an attention-based fusion technique. We measured the performance of each of the proposed architectures using various performance metrics. We conducted a rigorous error analysis of the classification performance. Our finding suggests that deep learning architectures trained only on text or speech data could underperform compared to architectures trained on both text and speech, thus, proving that multimodal systems can perform better than unimodal systems for emotion classification. Our future work involves extending our multimodal system to evaluate it on sentiment analyses and to include more modalities, such as video. We believe this is feasible while still using the same dataset because MELD contains both sentiment classes and videos. Additionally, our current architecture is designed to be simple and scalable, allowing us to easily extend it to include information learned from video contextualised embeddings.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[4] Z. Lian, B. Liu, and J. Tao, "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.

[5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," 2019.

[6] S. Dutta and S. Ganapathy, "Multimodal transformer with learnable frontend and self attention for emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6917–6921, 2022.

[7] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. Morency, "Memory fusion network for multi-view sequential learning," *CoRR*, vol. abs/1802.00927, 2018.

[8] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "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)*, (Melbourne, Australia), pp. 2236–2246, Association for Computational Linguistics, July 2018.

[9] S. Dutta and S. Ganapathy, "Hcam – hierarchical cross attention model for multi-modal emotion recognition," 2023.

[10] J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu, "HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*, (Barcelona, Spain (Online)), pp. 4190–4200, International Committee on Computational Linguistics, Dec. 2020.

[11] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. F. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," *CoRR*, vol. abs/1908.11540, 2019.

[12] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *AAAI Conference on Artificial Intelligence*, 2019.

[13] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," *CoRR*, vol. abs/2104.01778, 2021.

[14] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *CoRR*, vol. abs/1810.02508, 2018.

[15] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. F. Gelbukh, and E. Cambria, "Dialoguernn: An attentive RNN for emotion detection in conversations," *CoRR*, vol. abs/1811.00405, 2018.

[16] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[17] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[18] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, 2021.