Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Computer Science and Computational Intelligence 2025 (ICCSCI 2025)

Leveraging Pre-Trained Self-Supervised Learning Models for Low-Resource Multimodal Text and Speech Sentiment Analysis

Supandi[a], Eifel Linggar Navara[a], Alexander Agung Santoso Gunawan[a,1], Rilo Chandra Pradana[b]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[b]*Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Multimodal sentiment analysis (MSA) is designed to recognize human sentiment from multiple data modalities, offering richer context and improved accuracy compared to unimodal approaches. The MELD dataset provides text, audio, and visual modalities, which are extracted from episodes of the TV series Friends, with each utterance annotated with both emotion and sentiment labels. This study only utilizes text and speech modalities from the MELD dataset due to the visual features containing a significant amount of background information, which introduced irrelevant noise unrelated to the emotional states of the characters. This approach is particularly well-suited for podcast analysis, as podcast primarily rely on spoken language without consistent visual cues. However, the MELD dataset presents several challenges. These consist of loud audio, feature overlap between sentiment classes, and unbalanced data where neutral sentiments are dominate. To address these issues, a lightweight yet effective model is proposed that integrates BERT-based textual embeddings and Wav2Vec2 for audio features, employing an element-wise fusion strategy followed by a Multi-Head Attention and LSTM (MHA-LSTM) for contextual sequence modeling, which improves representation learning for sequence classification. The MHA-LSTM focuses on relevant input sequences, capturing long-term temporal dependencies. This model achieves competitive performance on MELD, with an overall accuracy of 67.66% and a weighted average F1-score of 0.65, which demonstrate that our approach offers strong accuracy and computational efficiency, making this model suitable for real-world applications where only text and audio modalities are being used.

Keywords: Multimodal Sentiment Analysis; MELD; BERT; Wav2Vec2; MHA-LSTM; Late Fusion.

[1]* Corresponding author. Tel.: +62-21- 5345830
E-mail address: aagung@binus.edu

1. Introduction

In recent times, a vast amount of content and many products have been widely consumed by the public. This leads to extensive public feedback and opinions for crucial product evaluation and development. This is where sentiment analysis is very helpful in this situation. Sentiment analysis is a method for opinion mining that is used to extract and analyze public reviews [1]. However, human reviews and opinions are expressed not just in text form but can be in several forms, such as speech and visual modalities [2, 3].

The combined insight of multiple modalities allows a better way to capture more complex human emotions and a better understanding of how humans feel. Therefore, multimodal sentiment analysis (MSA) nowadays has become increasingly important and popular in recent studies [4, 5]. MSA has greatly improved in public opinion analysis, education, healthcare, human-computer interaction, and recommendation systems by achieving comprehensive and accurate sentiment perception. This is why it's important to comprehend multimodal interaction development of better agents modeled after the human brain [6]. However, there are several difficulties in MSA, including several dataset classes that are often imbalanced and insufficiently large. Furthermore, each modality of real-world data can sometimes be noisy and lack information [7]. These difficulties result in limiting the effectiveness of fusion methods and deep learning models. Many prior works that depend only on unimodality, such as text or speech, result in constrained outcomes. Therefore, the model needs advanced feature extraction and fusion systems for multimodal information variance [8].

In this work, the MELD dataset is used as a benchmark for multimodal sentiment analysis, as it provides more than 1,400 multi-party dialogues and over 13,000 utterances from the TV series Friends [9]. The size of the MELD dataset is relatively small and has limited labeled data; this indicates that it is still considered a low resource for deep learning models [2]. Furthermore, it offers more realistic conversational flow, with multiple speakers and environmental noise; this provides more challenge in evaluating multimodal fusion and the pre-trained self-supervised model approach. Recent advanced studies, such as multi-level conflict-aware networks, as proposed by Gao et al., explicitly resolve the contradictions between unimodal and bimodal networks at two different (main and conflict) branches. This makes the model more robust in predictions while capturing the conflicting sentiment signals [8]. Similarly, shared-private memory networks, highlighted in recent studies, separate modality-specific private information from cross-modal data that are shared to learn shared representation and interaction of multimodal and unimodal data. Then there is the cross-modal attention technique; this architecture is particularly effective for capturing the integration and interdependencies between modalities [10]. This method has been proposed by Deng et al., demonstrating how a pair of multi-head attention can reflect the interactions on each modality that can contribute, based on their significance, to sentiment predictions. This work achieved 82.04% accuracy on other benchmark datasets in the CMU-MOSEI dataset. While these approaches show promising results, these studies often require complex architectures and substantial computational resources, making them less suitable for low-resource settings like the MELD dataset. Furthermore, advanced fusion methods focused on text-image analysis [11] rather than text-speech combination; this results in leaving the gap for efficient multimodal fusion in acoustic-based sentiment analysis.

Instead of conducting a complex deep learning model architecture, this work uses transformer-based models such as BERT [12, 13] to extract text modality and Wav2Vec 2.0 [14] for speech modality. Both typical models were already trained in a self-supervised environment on a large dataset. Thus, fine-tuning these models allows for the extraction of rich, generalizable features of text and speech modalities [12]; it also lowers the risk of overfitting and enables better generalization from small datasets like MELD. Then, to combine each modality embedding requires fusion strategies. Some existing fusion strategies in recent studies either depend on complex early fusion (concatenating features before classification) architecture that requires large datasets and careful alignment of each different data modality or hybrid fusion strategies that combine early and late fusion strategies for richer multimodal representations [3, 15]. And this technique generally achieves the best performance but increases model complexity and training requirements. Thus, to address this limitation, our approach is to implement a lightweight yet effective multimodal sentiment classifier that leverages pre-trained self-supervised models with a simple but robust late fusion approach [9]. The method also has been proven to be time-saving and practical for computational efficiency [16]. Our work emphasizes the usage of BERT and Wav2Vec2 to extract text and speech embeddings [5], followed by element-wise addition as the late fusion technique [4, 9]. The late fusion technique offers modularity and robustness to missing or noisy modalities. Then this fused representation is then processed through a multi-head attention layer and long short-term memory (MHA-LSTM) [4] for sequence-level modeling before final classification. To the best of our knowledge, this implementation of element-wise addition techniques with MHA-LSTM has not been explored in previous works. This work provides A simple, interpretable, and computationally efficient alternative is presented, demonstrating the effectiveness of element-wise addition for fusion, followed by the application of an attention mechanism to the fused feature (MHA-LSTM) [4] that still can capture cross-modal interactions and temporal dependencies without visual modality [10, 17].

1. Related Works

Representation learning in MSA has evolved from just putting features together to using more sophisticated methods that aim to understand information that is the same across different types and information that is unique to each type. Recent studies have demonstrated implementing multiple modalities improved significant sentiment classification robustness and accuracy. For instance, Shah et al. shows the effectiveness of bimodal sentiment analysis, showing that combining acoustic and linguistic features yields better sentiment classification than unimodal approaches [17].

A main challenge in MSA is the effectiveness of fusion techniques that are used to combine different modalities. There are already some fusion techniques that have been explored in past studies, such as early fusion, late fusion, and hybrid fusion. Early fusion concatenates features at the input level, while late fusion combines outputs at the decision level, and hybrid fusion is a method of combining late and early fusion approaches [3, 15]. For instance, Zhao et al. performs a multi-level fusion that shows the combination of early and late fusion models, enhanced by multi-granularity features, leads to an improvement of 1.3% in IEMOCAP datasets [3]. However, these strategies could hinder scalability to larger datasets or real-time applications, and the effectiveness of multi-granularity and fusion strategies often depends on the availability of high-quality and aligned multimodal data. This limitation is particularly relevant for datasets like MELD, where alignment and data quality issues may affect the effectiveness of advanced fusion strategies. Therefore, these fusion techniques present challenges such as data alignment, feature extraction, and increased model complexity.

In recent studies, the use of pre-trained models has been a popular solution to handle data scarcity and generalization, especially in low-resource settings. Pre-trained models such as BERT and RoBERTa have proven effective for extraction from contextualized text modality [2], while Wav2Vec2 has also proven effective for robust feature extraction from speech signals. For example, a recent multimodal system integrated RoBERTa for text, Wav2Vec2 for audio [17], and visual modalities, demonstrating good performance on the MELD dataset compared to unimodal architecture. Farhadipour et al. shows that transfer learning by fine-tuning pre-trained models on the target dataset and using feature concatenation as the fusion technique achieved 72.15% accuracy for sentiment analysis [2].

Despite these advances in MSA, many studies have several challenges, such as the need for large labeled datasets, the complexity of aligning multimodal data, and the computational demands of complicated fusion architectures. There remains a gap in conducting lightweight yet effective models that are robust and can generalize well in low-resource datasets such as the dataset used for this work. Thus, solving these challenges is essential for practical deployment of MSA in real-life applications.

2. Methodology

1.1. Dataset

This study utilizes the Multimodal Emotion Lines Dataset (MELD) [9]. This dataset is a publicly available dataset with 13,000 rows of utterances. Each utterance in the dataset is annotated with seven emotion and three sentiment labels [2] that are established for multimodal emotion recognition and sentiment analysis. The MELD datasets contain dialogues from the *Friends* TV series, annotated with both textual and audio information [9]. Table 1 shows that the dataset has already been partitioned with a ratio of approximately 75:10:15. The training set contains 9989 rows, the validation set contains 1109 rows, and finally, the test set contains 2610 rows.

Table 1. MELD dataset splitting with subset of classes.

Classes	Training	Validation	Test

Positive	2334	233	521
Neutral	4710	470	1256
Negative	2945	406	833

1.2. *Proposed Method*

The method of this study is to perform data collection from the MELD dataset [9] that has already been used multiple times by previous studies. The architecture consists of three main components, such as modality-specific feature extraction (BERT and Wav2Vec2) [11, 12, 18] and late fusion via element-wise addition and sequence-level modeling with Multi-head Attention and LSTM (MHA-LSTM) [4].

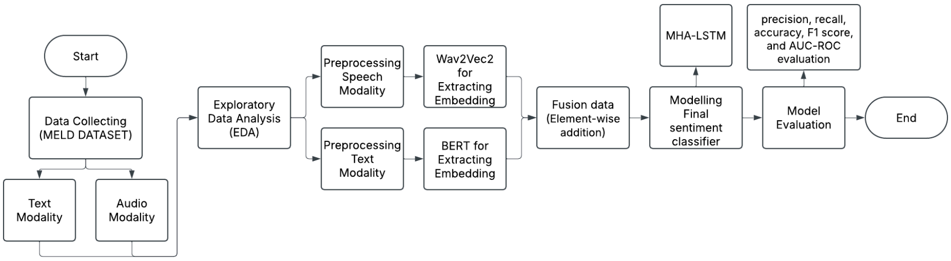
Start
Data Collecting (MELD DATASET)
Text Modality
Audio Modality
Exploratory Data Analysis (EDA)
Preprocessing Speech Modality
Preprocessing Text Modality
Wav2Vec2 for Extracting Embedding
BERT for Extracting Embedding
Fusion data (Element-wise addition)
MHA-LSTM
Modelling Final sentiment classifier
precision, recall, accuracy, F1 score, and AUC-ROC evaluation
Model Evaluation
End

Fig. 1. Research workflow

1.3. *Exploratory Data Analysis*

Conducting data analysis is an important step to properly prepare the dataset before feature extraction and training [9]. This study analyzes the data distribution, identifies missing values across each modality, and assesses the availability of each modality. Also, check audio availability by comparing dialogue and utterance IDs against the corresponding WAV files. When audio files were missing, quiet waveform placeholders were used to maintain consistent input dimensions. Table 2 presents the MELD dataset statistics, which show several challenges in the dataset, such as class imbalance, variability in utterance length, and modality-specific limitations, particularly within the audio modality. The variability in utterance length highlights the complexity of the dataset, as the model must effectively handle both brief and extended conversational turns.

Table 2. MELD data statistics.

MELD statistic	**Train**	**Dev**	**Test**
Unique words	5284	1425	2403
Average length	8.61	8.6	8.89
Max length	72	42	50
Number of dialogues	1038	114	280
Dyadic dialogues	463	52.0	128
Number of utterances	9989	1109	2610
Number of speakers	260	47	100
Utterance per dialogue	9.62	9.73	9.32
Emotions per dialogue	3.3	3.35	9.32
Average speakers per dialogue	2.71	3.01	2.66
Max speakers per dialogue	9	8	8
Emotion shift	5371	607	1353
Average duration	3.14	3.12	3.31

1.4. *Data Preprocessing*

Each utterance of raw text and audio is converted into two modality-specific forms during the preprocessing stage. Our approach to handling the text modality is to tokenize the raw utterance using the BERT tokenizer [18], with all sequences padded or truncated to a fixed length of 128 tokens. This process produces input_ids (integer token sequences) and corresponding attention_mask tensors, where the attention mask indicates which tokens in the input_ids should be attended to by the model. For the audio modality, raw waveform data are loaded as one-dimensional float tensors sampled at 16 kHz [14]. To handle stereo audio, the waveform is first resampled, if necessary, then processed and converted to mono by averaging across the channels. Then the waveform $\mathbf{x}$=[$x1$,$x2$,…,xN] is normalized to have values between -1 and 1 by dividing the maximum absolute value max(|$\mathbf{x}$|) among all samples and returning it as raw waveform tensors.

1.5. Multimodal model architecture

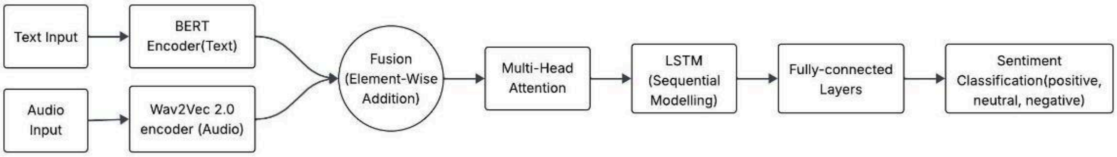
Text Input
BERT Encoder(Text)
Audio Input
Wav2Vec 2.0 encoder (Audio)
Fusion (Element-Wise Addition)
Multi-Head Attention
LSTM (Sequential Modelling)
Fully-connected Layers
Sentiment Classification(positive, neutral, negative)

Fig. 2. model architecture

1.1.1. BERT

The proposed model architecture integrates both textual and audio modalities to perform sentiment classification through a series of specialized encoding, fusion, and sequential modeling components. BERT (Bidirectional Encoder Representations from Transformers) [18] is a pre-trained language model that generates contextualized embeddings for input text by processing it bidirectionally, capturing semantic and syntactic information from both preceding and following words [13]; its output is a dense vector representation encoding the utterance's meaning.

1.1.2. Wav2Vec2

Wav2Vec 2.0 [14], a self-supervised model, encodes the voice audio using a multilayer Convolutional Neural Network (CNN). Following encoding, the resulting latent speech representation is subjected to masking before being input into a transformer network to create a contextualized representation of the speech audio. Both text and audio embeddings are fused via element-wise addition, creating a unified multimodal representation.

1.1.3. Multi-head Attention

MHA mechanism [4] enhances the model's ability to capture different parts of this fused representation in parallel, interdependence, and relationships within the combined features. Instead of computing a single attention distribution, MHA projects the input features into multiple subspaces, which in this architecture uses 4 head with learnable linear transformations for queries (Q), keys (K), and values (V) matrices [3, 15]. To obtain attention scores for each head, the model takes the dot product of queries and keys, scaling the square root of the key dimension, applying softmax to obtain weights, and weighing the values accordingly [15]. These weights are then used to aggregate the value vector. Each head is concatenated to produce input features focusing on aspects or patterns within the sequence, such as syntactic structure, semantic meaning, or positional relationships. This allows the model to simultaneously pay attention to multiple information sources.

1.1.4. LSTM

Subsequently, an LSTM [6] network models the sequential nature of dialogue by processing the sequence of fused embeddings. The core of the LSTM architecture consists of the cell state that carrying accumulated information over many time steps for long-term memory and the hidden state, which consists of 512 hidden sizes that represents the output of LSTM cell at each time step and capture the current context of short-term information

[12]. This model updated each cell state at each time step through gating mechanism such as, forget gate, input gate, and output gate [12]. The forget gate filters out irrelevant information from the previous cell state. Input gate determines the new information should be added to the cell state. Additionally, the output gate, which serves as both short-term memory and output, controls which portion of the memory affects the current hidden state [12].

1.1.5. Fully connected layers

Fully connected layers [15] serve as the final classification layer that converts the last time step hidden state from LSTM, which is fixed-length vector of 512-dimensions into logits, facilitating supervised learning for classification tasks. In FC layers each neuron is connected to every element of the input vector, allowing the model to combine all learned features globally and produce a final prediction for each class, which predicts the sentiment label (positive, neutral, or negative) [15].

1.6. *Specific-Modality Feature Extraction*

This typical method ensures uniformity and mitigates variations in the datasets. The text and audio modalities are transformed into feature representations suitable for multimodal late fusion and final sentiment classification. The input_ids and attention_mask are extracted using a pre-trained BERT model, which generates contextualized word embeddings with 768 dimensions for each utterance [17]. For the audio data, a pre-trained Wav2Vec2 model is utilized to extract raw acoustic features [17]. The hidden states of the model are extracted to capture contextualized audio embeddings, each also with 768 dimensions, providing a suitable representation of the acoustic properties of the audio modalities. These feature extraction steps leverage self-supervised learning to generate feature embeddings from both text and audio modalities.

1.7. *Fusion Technique*

The text and audio sequences have to be aligned before they can be fused, so both modalities are truncated to the minimum sequence length across the batch. To effectively integrate textual and acoustic features, the two extracted features are combined, and element-wise addition is implemented to fuse these modality embeddings. Element-wise addition takes each element from the text vector and adds it to the corresponding element of the speech vector to form a new fused vector.

1.8. *Model Training*

The final multimodal sentiment classifier is designed using a fully connected neural network incorporating Multi-Head Attention and LSTM (MHA-LSTM) for effective sequential modeling [4]. The MHA enables the capture of both cross-modal and temporal dependencies in both features with 768 dimensions as an input, where it supports the limitations of simple fusion by focusing on the most relevant fused features for sentiment analysis. Then, the vectors are input into LSTM to capture sequential dependencies and context. The LSTM produces a 512-dimensional vector from the last time step, where this vector is then fed into a fully connected layer with softmax activation for three sentiment classes (positive, neutral, and negative) [10]. For the training settings, the Adam optimizer was used with a learning rate of 1e-5 and a batch size of 8, training up to 5 epochs. Training stops based on validation loss. Furthermore, dropout with a rate of 0.1 was applied after the fusion layer [3]. The training configurations are detailed in Table 3.

Table 3. Model hyperparameter settings.

Hyperparameter	**MELD (Our work)**
Optimizer	Adam
Batch size	8
Learning rate	1e-5
Epoch number	5
Dropout	Default in BERT/Wav2Vec2 (0.1)

1.9. *Model Evaluation*

To assess the MHA model architecture performance, the evaluation is examined with these main evaluation metrics, such as accuracy, F1-score, precision, and recall [19]. With accuracy as the primary metric used for evaluation, which measures the proportion of correct predictions out of precision and the overall correctness of the model. Then, there is precision, which quantifies the proportion of correctly predicted positive instances among all predicted as positive. To measure the proportion of correctly predicted positive instances among all actual positive instances, recall metrics were applied, followed by using weighted F1-score to address the imbalanced dataset by calculating the harmonic mean of precision and recall. These evaluation metrics reflect the model's performance across all classes proportionally to their frequency. Evaluation using a confusion matrix is also employed [20], which summarizes the performance of a classification model by showing the counts of correct and incorrect predictions broken down by each class. The diagonal elements represent the correct predictions for each class. And off-diagonal represents misclassifications, which classes are confused with each other.

3. **Result & Discussion**

The result of our MHA model, which leverages specific-modality features extraction. BERT for text, Wav2Vec2 for speech [17]. Then using element-wise fusion, and MHA-LSTM classifier [4] on the MELD dataset which is emotionally rich, and contextually grounded interactions [9]. Below is the summary of the primary result:

Table 4. Confusion matrix on sentiment classes.

	Positive	**Neutral**	**Negative**
Positive	348	104	69
Neutral	120	964	172
Negative	174	205	454

Table 5. Model performance on sentiment classes.

Classes	**Weighted F1-score**	**Recall**	**Precision**
Positive	0.59	0.66	0.54
Neutral	0.76	0.76	0.75
Negative	0.59	0.54	0.65

Table 5 presents the model's performance for each sentiment class, with additional confusion matrix details shown in Table 4. These results indicate that our model performs best on the neutral class, achieving an F1 score and recall of 0.76 and a precision of 0.75. With a solid recall of 0.66 and a precision of 0.54, the positive class achieves an F1 score of 0.59. This finding implies that our model is able to capture most positive instances but misclassify other classes as positive, leading to capturing false positives (FP). This can be reflected in Table 4, where 104 positive samples are misclassified as neutral and 69 as negative samples. For the negative class, our model achieves an F1 score of 0.59 but with higher precision of 0.65. This shows that our model is more capable of predicting negative sentiment but is missing a larger portion of the actual negative class. This is also represented in the confusion matrix on Table 3, where 205 negative samples are misclassified as neutral and 69 as negative samples. Based on our observation, the performance on positive and negative sentiment may be due to class imbalance, which the neutral sentiment class tends to dominate in the MELD dataset, or overlapping features between these classes.

Table 6. Model performance on text, audio, and text + audio with MELD dataset.

Model	**Modality**	**Accuracy (%)**	**Weighted F1-score**	**Recall**	**Precision**
BERT	Text	67.43	0.65	0.64	0.65
Wav2Vec2	Audio	48.12	0.21	0.33	0.16
BERT + Wav2Vec2 + MHA-LSTM (Ours)	Text + Audio	**67.66**	**0.65**	**0.66**	**0.65**

To assess the performance of the text-only and audio-only models in comparison to the multimodal model, the models are evaluated, as shown in Table 6. The text-only model that used BERT already performed strongly with an accuracy of 67.43 and an F1 score of 0.65 on the MELD dataset; however, the audio-only model, which used Wav2Vec2, performed significantly worse with an accuracy of 48.12% and an F1 score of 0.21. These results show the multimodal model only improves a little bit over the unimodal baseline. This slight improvement might result from the fact that textual information carries the most explicit and direct cues for sentiment, while some utterances are ambiguous, making audio less influential, or the element-wise fusion and MHA-LSTM architecture are not fully exploring the complementary information between modalities. Therefore, incorporating audio features can further enhance the model's ability to capture subtle cues and use more sophisticated fusion techniques, leading to more robust sentiment classification.

Table 7. Comparing model performance based on previous literature on MELD dataset.

Model Framework	**Modality**	**Accuracy (%)**	**Weighted F1-score**	**Recall**	**Precisio n**
Cross-Modal Multi-head Attention [10]	Text + Audio	60.74	0.55	-	-
RoBERTa last four hidden layers + Wav2Vec2 [17]	Text + Audio	**68.4**	0.68	**0.68**	0.68
BERT + Wav2Vec2 + MHA-LSTM (Ours)	Text + Audio	67.66	**0.67**	0.67	**0.68**

Table 7 provides a comparison of various MHA frameworks utilizing both text and audio modalities on the MELD dataset. The proposed BERT + Wav2Vec2 + MHA-LSTM model achieves an accuracy of 67.66% and a weighted F1-score of 0.67, demonstrating competitive performance relative to existing approaches. While the RoBERTa last four hidden layers + Wav2Vec2 model attains the highest accuracy at 68.4% and a slightly higher weighted F1-score of 0.68 [17]. Several architectural choices determine the performance improvement of their work, where they utilize RoBERTa, as a transformer-based encoder pretrained on a larger corpus than BERT, to capture richer contextual embedding in text modality. Unlike our model, which uses element-wise fusion followed by LSTM, their approach performs late fusion through direct concatenation of a large feature vector with 3840 dimensions and feeds it into a deep, fully connected network with five linear layers. The LSTM module in our architecture explicitly handles temporal tracking, which is advantageous when modeling sequential dependencies across utterances, whereas their model favors a stronger encoder-level representation. Additionally, the resource-intensive training with early stopping patience of 30 epochs and higher dimensional embeddings is not feasible in low-resource computation. In contrast, our model performs temporal interpretability and computational efficiency but with slightly worse performance for better sequence modeling alignment with the MELD dataset.

The proposed model closely matches these results, outperforming the Cross-Modal Multi-Head Attention framework, which records an accuracy of 60.74% and a weighted F1-score of 0.55 [10]. This is due to their model using basic word embeddings and simple MFCC audio features, while this study utilizes BERT and Wav2Vec2, which give more meaningful features from both text and speech. Also, their model struggles to capture the flow of conversation effectively due to the absence of sequence-based layers. In contrast, our model incorporates LSTM, enabling it to better understand the progression of emotions throughout the dialogue.

4. Conclusion

This study proposed a lightweight and effective multimodal sentiment analysis model that integrates BERT-based textual embeddings with Wav2Vec2 audio features using an element-wise fusion strategy, followed by a Multi-Head Attention and LSTM (MHA-LSTM) module for contextual sequence modeling. Which is conducted on the MELD dataset. The experimental findings show that the model is better at classifying neutral sentiment, where the F1-score is 0.76. Furthermore, the multimodal model performance is slightly better than the unimodal baselines that just apply text or speech modalities, with a weighted F1-score of 0.65 and overall accuracy of 67.66%. The additional audio modality as an input improves the model's capacity to capture complex sentiment expression, even though text modality is still the main contributor to the model. The slight but steady improvement demonstrates the late fusion strategy and sequence modeling technique on MSA.

Overall, the results demonstrate that our proposed MHA-LSTM-based fusion model strikes a good balance between performance and efficiency for multimodal sentiment analysis on MELD. It achieves competitive accuracy and F1-score using only text and audio, highlighting the value of effective fusion and contextual modeling in conversational sentiment understanding. Future work will focus on enhancing visual feature extraction by concentrating on the facial expression regions of individuals and eliminating irrelevant information, thereby reducing noise and enabling more accurate experiments and analyses. Additionally, utilizing RoBERTa as the text encoder with larger corpus, which may further improve the performance of multimodal sentiment analysis.

5. Author Contribution

The study was initiated and planned by AASG. The experiments were conducted by ELN. SS provided guidance throughout the writing process. All authors have reviewed and approved the final manuscript.

6. Data and Materials Availability

This study publicly shares the code, deep learning model, and application on GitHub. You can access them at https://github.com/EifelLN/MELD-msa.

References

[1] Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A. Multimodal Sentiment analysis: a Systematic Review of history, datasets, Multimodal Fusion methods, applications, Challenges and Future Directions. Information Fusion 2023;91:424–44. https://doi.org/10.1016/j.inffus.2022.09.025.

[2] Farhadipour A, Ranjbar H, Chapariniya M, Vukovic T, Ebling S, Dellwo V. Multimodal Emotion Recognition and Sentiment Analysis in Multi-Party Conversation Contexts. ArXivorg 2025. https://arxiv.org/abs/2503.06805.

[3] Cai Y, Li X, Zhang Y, Li J, Zhu F, Rao L. Multimodal Sentiment Analysis Based on multi-layer Feature Fusion and multi-task Learning. Scientific Reports 2025;15. https://doi.org/10.1038/s41598-025-85859-6.

[4] Banerjee D, Lygerakis F, Makedon F. Sequential Late Fusion Technique for Multi-modal Sentiment Analysis. Proceedings of the 14th Pervasive Technologies Related to Assistive Environments Conference 2021:264–5. https://doi.org/10.1145/3453892.3461009.

[5] Aliyu Y, Sarlan A, Usman Danyaro K, Rahman ASBA, Abdullahi M. Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. IEEE Access 2024;12:66883–909. https://doi.org/10.1109/access.2024.3398635.

[6] Paraskevopoulos G, Georgiou E, Potamianos A. MMLatch: Bottom-up Top-down Fusion for Multimodal Sentiment Analysis. ArXivorg 2022. http://arxiv.org/abs/2201.09828.

[7] Zhao X, Chen Y, Liu S, Tang B. Shared-Private Memory Networks for Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing 2022;14:1–13. https://doi.org/10.1109/taffc.2022.3222023.

[8] Gao Y, Wu H, Zhang L. Multi-level Conflict-Aware Network for Multi-modal Sentiment Analysis. ArXivorg 2025. http://arxiv.org/abs/2502.09675.

[9] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: a Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. ArXivorg 2018. http://arxiv.org/abs/1810.02508.

[10] Deng L, Liu B, Li Z. Multimodal Sentiment Analysis Based on a Cross-Modal Multihead Attention Mechanism. Computers, Materials & Continua/Computers, Materials & Continua (Print) 2023;0:1–10. https://doi.org/10.32604/cmc.2023.042150.

[11] Zhao T, Li M, Chen K, Wang L, Zhou X, Chaturvedi K, et al. Enhancing Sentiment Analysis through Multimodal Fusion: A BERT-DINOv2 Approach. ArXivorg 2025. http://arxiv.org/abs/2503.07943.

[12] Jafarzadeh P, Rostami AM, Choobdar P. Speaker Emotion Recognition: Leveraging Self-Supervised Models for Feature Extraction Using Wav2Vec2 and HuBERT. ArXivorg 2024. http://arxiv.org/abs/2411.02964.

[13] Zou W, Ding J, Wang C. Utilizing BERT Intermediate Layers for Multimodal Sentiment Analysis. 2022 IEEE International Conference on Multimedia and Expo (ICME) 2022. https://doi.org/10.1109/icme52920.2022.9860014.

[14] Jain R, Barcovschi A, Yiwere M, Bigioi D, Corcoran P, Cucu H. A Wav2vec2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition. ArXivorg 2023. https://doi.org/10.48550/arXiv.2204.05419.

[15] Li J, Zhang Z, Lang J, Jiang Y, An L, Zou P, et al. Hybrid Multimodal Feature Extraction, Mining and Fusion for Sentiment Analysis. ArXivorg 2022. http://arxiv.org/abs/2208.03051.

[16] Wagner J, Andre E, Lingenfelser F, Kim J. Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. IEEE Transactions on Affective Computing 2011;2:206–18. https://doi.org/10.1109/t-affc.2011.12.

[17] Shah S, Ghomeshi H, Vakaj E, Cooper E, Mohammad R. An Ensemble-Learning-Based Technique for Bimodal Sentiment Analysis. Big Data and Cognitive Computing 2023;7:85. https://doi.org/10.3390/bdcc7020085.

[18] Zhao Z, Wang Y, Wang Y. Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition. ArXivorg 2022. https://arxiv.org/abs/2207.04697.

[19] Hu G, Lin T-E, Zhao Y, Lu G, Wu Y, Li Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. ArXivorg 2022. http://arxiv.org/abs/2211.11256.

[20] Dutta S, Ganapathy S. HCAM -- Hierarchical Cross Attention Model for Multi-modal Emotion Recognition. ArXivorg 2023. http://arxiv.org/abs/2304.06910.