ATFSC: Audio-Text Fusion for Sentiment Classification

Aicha Nouisser[1], Nouha Khediri[2][a], Monji Kherallah[3][b] and Faiza Charfi[3][c]
[1]*National School of Electronics and Telecommunications of Sfax, Tunisia*
[2]*Faculty of Computing and Information Technology, Northern Border University, Rafha, K.S.A.*
[3]*Faculty of Sciences of Sfax, University of Sfax, Tunisia*
ﬁ

Keywords: Sentiment Analysis, Bimodality, Transformer, BERT Model, Audio and Text, CNN.

Abstract: The diversity of human expressions and the complexity of emotions are specific challenges related to sentiment analysis from text and speech data. Models must consider not only text but also nuances of intonation and emotions expressed by voice. To address these challenges, we created a bimodal sentiment analysis model named **ATFSC**, that organizes emotions based on textual and audio information. It fuses textual and audio information from conversations, providing a more robust analysis of sentiments, whether negative, neutral, or positive. Key features include the use of transfer learning with a pre-trained BERT model for text processing, a CNN-based audio feature extractor for audio processing, and flexible preprocessing capabilities that support different dataset formats. An attention mechanism was employed to perform a bimodal fusion of audio and text features, which led to a notable performance optimization. As a result, we observed a performance amelioration in the accuracy values such as 64.61%, 69%, 72%, 81.36% on different datasets respectively IEMOCAP, SLUE, MELD, and CMU-MOSI.

1 INTRODUCTION

Due to growing demand and many unsolved problems, numerous studies focus on emotion recognition through visual, verbal, and nonverbal expressions. Exploring different modalities (video, audio, text) is essential, as each contributes differently to system reliability. Experimental tests are crucial in selecting the appropriate methods (Dvoynikova and Karpov, 2023). Deep learning algorithms have recently shown success in fields such as image classification, machine translation, speech recognition, and text recognition. This advancement led to research into human emotions and their representation through artificial intelligence, including emotional dialogue models (Yoon et al., 2018). For more details about the methods used for uni-modal emotion and sentiment recognition, the readers can refer to (Khediri et al., 2017; Khediri et al., 2022).

Emotion and sentiment recognition are essential to improve human-machine interactions. Despite advances in machine learning, machines struggle to distinguish human emotions adequately. Identifying emotions in speech enables automatic recognition of an individual's emotional state, focusing on audio features (Bhaskar et al., 2015). However, few approaches have focused on detecting emotions from text data. Text is a key communication method, extracted from sources such as books, newspapers, and web pages. Natural language processing techniques allow emotion extraction from textual input (Ye and Fan, 2014).

Sentiment analysis is widely used in various fields, providing insight into public emotions and opinions. Applications include customer feedback analysis, real-time social media monitoring, market research, brand reputation management, and political campaigns. It also plays a role in financial markets, healthcare, media, and academic research (Jim et al., 2024).

Automatic analysis systems are crucial for recognizing emotions across speech, text, and bimodal forms. This study presents and evaluates the bimodal approach **ATFSC** (**A**udio-**T**ext **F**usion for **S**entiment **C**lassification), which integrates a Bidirectional Encoder Representations from Transformers (BERT) model for text and a Convolutional Neural Network (CNN) based audio extractor.

The fusion method improves accuracy and robustness by combining audio and text data, as shown by

[a] https://orcid.org/0000-0002-0189-7986
[b] https://orcid.org/0000-0002-4549-1005
[c] https://orcid.org/0009-0003-9508-0831

Nouisser, A., Khediri, N., Kherallah, M. and Charfi, F.
ATFSC: Audio-Text Fusion for Sentiment Classification.
DOI: 10.5220/0013178300003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 750-757
ISBN: 978-989-758-737-5; ISSN: 2184-433X

experimental results on different datasets.

The goal of this article is to present an effective technique for recognizing bimodal feelings: Combining audio and text for feeling categorization.

The rest of the paper is organized as follows. Section 2 gives an overview of related methods. In Section 3, we present an overview of the techniques used in our model. The suggested model for identifying emotions and sentiment used audio and text modalities is outlined in Section 4. Section 5 presents the datasets used. Thereafter, we report the obtained results in Section 6. Finally, in Section 7, we conclude and outline future work.

2 STATE OF THE ART

In this section, we present a brief overview of previous works that focus on emotion and sentiments recognition using only two modalities (text and audio), which is the interest of this paper.

We found in literature, a multi-headed attention mechanism for bimodal sentiment analysis using audio and text modalities was proposed by (Deng et al., 2024), within a transformer model with cross-modality attention, achieving accuracies of 60.74% for 3-class classification and 55.13% for 7-class classification on the MELD dataset, with an accuracy of 82.04% for CMU-MOSEI.

A multitasking preprocessing and classification system is proposed by (Dvoynikova and Karpov, 2023) , using the EmotionHuBERT and RoBERTa models on the CMU-MOSEI database. Accuracy for sentiment recognition is 63.5% and for emotion recognition is 61.4%, measured using the macro F-score. For classification, the approach uses logistic regression, with the recognition of 3 classes of feelings and 6 classes of binary emotions. The means used in this research are audio recording, and text.

Furthermore, CM-BERT (Cross-Modal BERT), evaluated on CMU-MOSI with an accuracy of 44.7%, was proposed by (Voloshina and Makhnytkina, 2023). This model uses multimodal attentional masking to efficiently integrate textual and audio modalities in sentiment analysis. In addition, DialogueRNN for sentiment classification on MELD was introduced by (Poria et al., 2018), achieving a weighted average accuracy of 67.65%. This method uses intermediate fusion to integrate text and audio data.

Also, an icon based model for multimodal sentiment analysis, was proposed by (Sebastian et al., 2019) evaluated on MELD with a weighted average accuracy of 63.0%. This model uses dynamic cross-modality fusion to integrate audio and text data.

Furthermore, a multimodal sentiment analysis on a YouTube dataset was conducted by (Poria et al., 2016), achieving an accuracy of 66.4%. This method uses decision-level fusion to combine modalities text and audio . The emotion and sentiment identification bimodal systems aforementioned are briefly summarized in Table 1.

Table 1: Summary of Related Works of Bimodal Emotion and Sentiment Recognition.

Bimodal: Text and Speech				
Works	**Model**	**Dataset**	**Accuracy**	**Fusion**
(Deng et al., 2024)	Multi-headed attention (Transformer)	MELD, CMU-MOSEI	60.74% (3-class sentiment), 55.13% (7-class sentiment) for MELD, 82.04% for CMU-MOSEI	Cross-modality attention
(Dvoynikova and Karpov, 2023)	Emotion HuBERT + RoBERTa	CMU-MOSEI	63.5% (sentiment), 61.4% (emotion)	Early Concat + Late Multi-Head Attention
(Voloshina and Makhnytkina, 2023)	CM-BERT (Cross-Modal BERT)	CMU-MOSI	44.7%	Multi-modal attention masking
(Poria et al., 2018)	DialogueRNN (dRNN)	MELD	67.65% W-AVG	Intermediate Fusion
(Sebastian et al., 2019)	ICON (Icon-based Model for multimodal sentiment analysis	MELD	63.0% W-AVG	Inter-modality dynamic fusion
(Poria et al., 2016)	Multimodal sentiment Analysis	YouTube	66.4%	decision-level fusion

3 TECHNIQUES USED

3.1 Transformers

For a long time, reducing the sequential computational load has been a critical issue for NLP applications. Despite numerous suggested solutions, NLP remained dependent on linear or logarithmic dependency. Transformers offer a simpler structure, eliminating recurrent and convolutional layers, and adapting their architecture to allow a constant number of operations based on attention-weighted positions. BERT and GPT2 are the most popular transformer-based models.

In this context, we chose BERT, a powerful method for extracting textual representations due to its ability to capture bidirectional word context. BERT (Lee and Toutanova, 2018) is suitable for various neurolinguistic tasks. The next section will explore BERT's architecture, its operation, and its use in optimizing performance for text classification and other machine learning applications.

3.2 BERT: Bidirectional Encoder Representations from Transformers

For the text component of our system, we suggest using the BERT BASE model. BERT uses a multilayer, bidirectional Transformer encoder to capture the context of words. The model undergoes two key stages: pre-training and fine-tuning.

During pre-training, BERT learns from unlabeled data through tasks like masked language modeling (MLM) and next-word prediction. In the fine-tuning phase, BERT is adjusted using labeled data for specific tasks(Lee and Toutanova, 2018).

BERT excels at predicting hidden words by considering their surrounding context, enabling a two-way learning process. It is available in two main sizes: the base model and the large model.

BERT BASE is composed of 12 layers, with a hidden size of 768, and uses 12 self-attentive heads, totaling 110 million parameters. BERT LARGE has 24 layers, a hidden size of 1024, and 16 auto-attention heads, with a total of 340 million parameters. Thanks to these various dimensions, users can opt for a model tailored to their particular requirements in terms of performance and digital resources.

3.3 Convolutional Neural Networks CNN

CNN is a popular deep learning method that learns directly from input, without requiring feature extraction. An example with multiple convolutions and pooling layers. CNN improves the design of classical ANNs like MLP networks by optimizing parameters at each layer for meaningful outputs and reducing model complexity. Dropout in CNNs helps address overfitting issues typical in traditional networks (Sarker, 2021).

Convolutional neural networks are designed to handle different two-dimensional shapes and are therefore commonly used in the fields of visual recognition, medical image analysis, image segmentation, natural language processing and many others. The ability to automatically discover essential features from the input without the need for human intervention makes it more powerful than a traditional network.

4 PROPOSED MODEL: ATFSC

Since a single modality can result in unreliable emotion recognition, our system integrates both audio and textual information for better emotional state capture. This approach, named Bimodal Sentiment Recognition: Audio-Text Fusion for Sentiment Classification (**ATFSC**), as shown in Figure 1, allows for more accurate and nuanced sentiment analysis.

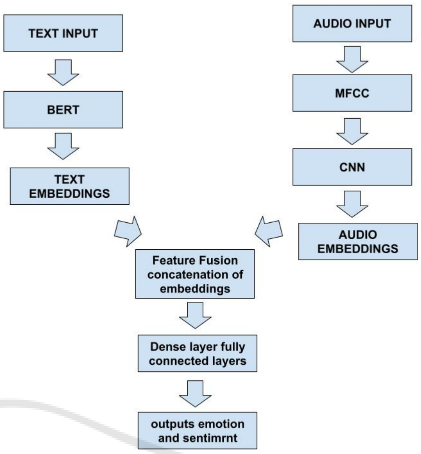
TEXT INPUT
BERT
TEXT EMBEDDINGS
AUDIO INPUT
MFCC
CNN
AUDIO EMBEDDINGS
Feature Fusion concatenation of embeddings
Dense layer fully connected layers
outputs emotion and sentimrnt

Figure 1: ATFSC Architecture.

By combining verbal and non-verbal cues, the model captures the complexity of emotions that a single modality may not fully address. The three main elements of the architecture are a text-processing module, an audio-processing module and a bimodal fusion module.

4.1 Audio Processing Module

In this module, we used mel-frequency cepstral coefficients (MFCC) to analyze audio characteristics, facilitating emotion recognition through tone, rhythm, and intonation. These sound signals are crucial for translating feelings. At the same time, tokenization was employed to process textual information by breaking the text into tokens, allowing the model to capture and analyze feelings and sentiments through digital representations.

4.2 Text Processing Module

The text processing module uses BERT to capture linguistic nuances with contextual embeddings. BERT weights are retained during training to preserve pre-trained knowledge. In the audio processing module, a customized CNN feature extractor with 2D convolutions, batch normalization, and ReLU activation is used to extract sound feature vectors.

4.3 Bimodal Fusion Module

Next, text and audio representations are merged using weighted attentive fusion with learnable weights, capturing their complementarities. Custom layer normalization (BertLayerNorm) stabilizes learning. The merged properties are processed through a final linear layer with a softmax function to obtain emotion probabilities (positive, negative, neutral). This bimodal approach enhances sentiment analysis by leveraging both verbal and vocal data. The model is optimized for the different datasets used and it offers a nuanced and precise understanding of human emotions.

5 DATASET USED

5.1 IEMOCAP

The IEMOCAP (Interactive Emotional Dyadic Motion Capture Database)(Firdaus et al., 2020) database gathers videos of didactic interactions between two pairs of 10 speakers, divided into 10 hours of dialogues and categorized according to very specific emotions (anger, excitement, joy, frustration, neutrality and sadness). It also includes continuous properties such as valence, activation and dominance.

5.2 MELD (A Multimodal Emotion Recognition)

MELD (Poria et al., 2018), also known as EmotionLines Multimodal, advances sentiment detection in discussions with 13,000 utterances from 1,433 “Friends” conversations. The corpus includes audio, visual, and textual formats, promoting a more effective understanding of emotions (Khediri et al., 2024).

According to earlier studies (Khediri et al., 2024), emotion analysis in MELD is difficult since each interaction often contains multiple speakers but few utterances.

5.3 SLUE

SLUE (Shon et al., 2022) provides a benchmark for examining pipelined methods and end-to-end strategies, from speech to labeling. It encourages research in oral language comprehension with a shared evaluation framework, basic models, and an open-source kit for replication.

The SLUE benchmark includes two datasets: SLUE-VoxPopuli and SLUE-VoxCeleb. SLUE-VoxPopuli contains nearly 5,000 speech recordings totaling 14.5 hours, covering training, verification, and testing sequences.

5.4 CMU-MOSI

The CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSI) dataset: This English-language dataset includes audio, text, and video formats aggregated from 2,199 annotated video segments collected from monologue movie reviews on YouTube. It proposes a specific method to analyze emotion recognition in movie reviews (Wu et al., 2024).

6 RESULTS AND DISCUSSION

6.1 Expriments 1 of ATFSC

Our model **ATFSC** was tested on the IEMOCAP dataset as first experiment. The Table 2 shows that the accuracy achieved was 64.61% using an attention mechanism.

Table 2: Performance of Our Model on IEMOCAP.

Model	**Dataset**	**Accuracy**	**Fusion**
Our model	IEMOCAP	64.61%	Attention mechanism

As shown in Figure 2, the graph shows the model’s accuracy evolution. The validation curve (orange) remains slightly higher than the training curve (blue), both stabilizing around 0.64 for validation and 0.63 for training after 2-3 epochs.

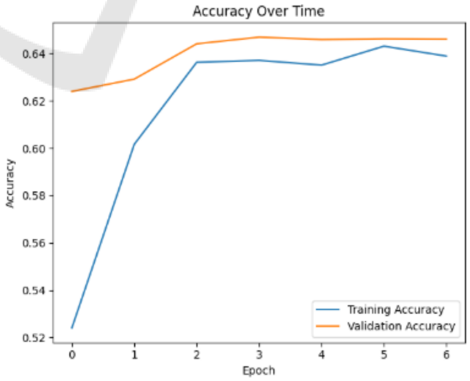

Figure 2: Training and Validation Accuracy Graph of Experiment 1.

As shown in Figure 3, the confusion matrix of a three-class sentiment classification model reveals poor performance. The negative, neutral, and positive classes have correct classification rates of 26.7%, 36.1%, and 20.6%, respectively. True negatives are often misclassified as neutral or positive, while the majority of true positives are classified as negative, indicating a model bias towards the negative class.

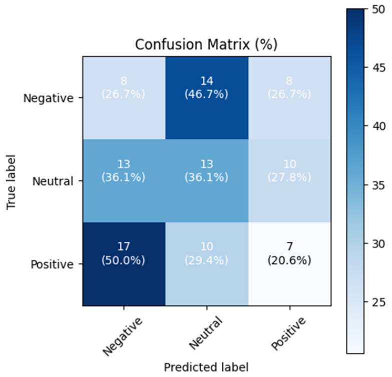

Figure 3: Confusion Matrix of a Classification Model of Experiment 1.

6.2 Experiments 2 of ATFSC

A second experiment was conducted to analyze the results from different evaluation metrics on the MELD and SLUE datasets. During the training session, as shown in Table 3, the loss is gradually reduced, from 1.0159 to 0.9186, indicating continued learning progress on the exercise data. At the same time, training accuracy increases slightly, from 0.5664 to 0.5799.

Table 3: Training and Validation Values of Experiment 2.

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
0	1.0159	0.5664	0.8692	0.6872
1	0.9810	0.5590	0.8618	0.6747
2	0.9621	0.5585	0.8250	0.6868
3	0.9375	0.5712	0.7995	0.6900
4	0.9287	0.5737	0.7957	0.6904
5	0.9241	0.5756	0.7941	0.6921
6	0.9186	0.5799	0.7924	0.6927

Furthermore, the validation loss decreases from 0.8692 to 0.7924, indicating that the model is increasingly able to be generalized to validated data. Finally, validation accuracy also increases, from 0.6872 to 0.6927. This suggests an optimization of predictive performance on the same information.

The diagram shown in Figure 4, illustrates how training accuracy and validation accuracy have progressed over the different periods. The blue curve (training acc) illustrates the accuracy of the training information: It starts at around 56.7%.

It undergoes a slight decrease until epoch 2, then it gradually increases to reach around 58% during epoch 6. The overall progression is rather modest (+1.3%). The orange curve (val acc) illustrates the accuracy of the validation information: It starts higher, around 68.7%. It undergoes a slight decrease until epoch 1. Then, it gradually increases to reach around 0.69% at epoch 6. In conclusion, this graph shows that the model is progressing satisfactorily, reducing losses and improving the accuracy of both training and validation data. This indicates an effective learning process and good generalization potential.

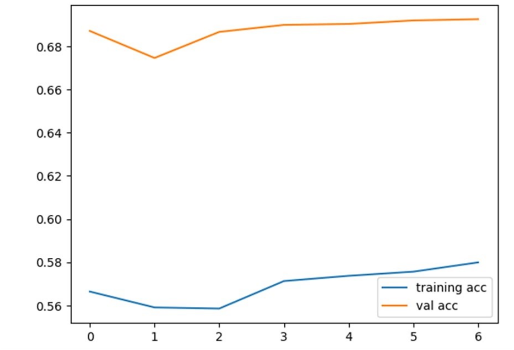

Figure 4: Training and Validation Accuracy Graph of Experiment 2.

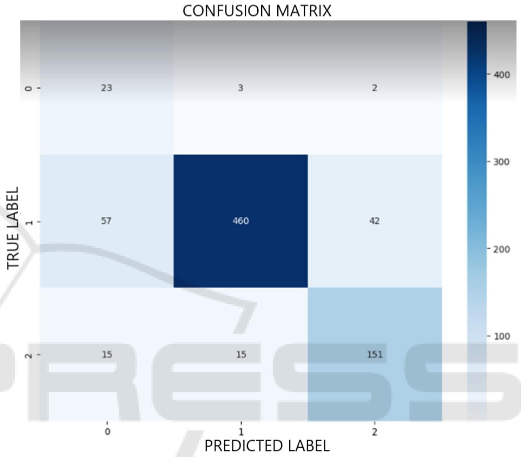

Figure 5: Confusion Matrix of a Classification Model of Experiment 2.

The confusion matrix of our first experiment is illustrated in Figure 5, which establish three classes for categorizing feelings in our model: Class 0 corresponds to negative, class 1 to neutral, and class 2 to positive.

The rows of the confusion matrix represent the true labels and the columns represent the model predictions.

Our results show that negative class was correctly classified, while neutral and positive classes were misclassified. where 57 cases of neutral class were misclassified as negative and 42 as positive class.

However, in positive class, 15 cases were misclassified as neutral. 151 cases were correctly classified, while 15 were misclassified as negative and neutral.

6.3 Expriments 3 of ATFSC

To deepen our analysis, a third experiment was conducted on the MELD and SLUE dataset. This phase will give us a better understanding of the optimizations performed on our **ATFSC** model. For the amelioration of this latter, we chose to modify the hyperparameters by increasing the rate of knowledge acquisition in 1e-5. This modification aims to improve the result, maintain the convergence of the model in place, and highlighted the need to improve the hyperparameters in the machine learning process.

Table 4: Training and Validation Values of Experiment 3.

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
0	37.85	0.4135	48.12	0.2282
1	83.81	0.4762	63.44	0.7146
2	134.06	0.6281	74.37	0.7046
3	124.33	0.6234	57.36	0.7245
4	108.02	0.6406	55.09	0.7269
5	104.96	0.6421	53.08	0.7279
6	103.53	0.6416	52.91	0.7276

Reagarding the Table 4 below, it can be seen that the training loss starts at 37.85 and increases significantly to 134.05 at epoch 2. The training accuracy gradually increases from 0.4135 to 0.6416. For the verification data, the validation loss decreases from 48.12% to 52.91%, while the validation accuracy improves significantly from 0.2282 to 0.7276, at the end of training.

These favorable developments on the training and validation indicators suggest that the model is making significant progress to the regulation of the learning rate. It appears that the model is better able to assimilate the specificities of training data, while generalizing more effectively to validated data.

The training and validation accuracy graph of the second experiments of our **ATFSC** is illustrated in Figure 6. The blue curve (training acc) illustrates the accuracy of the training information: It starts around 41 %, then stabilizes around 64%.

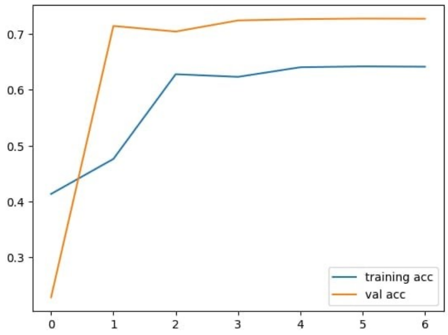

Figure 6: Training and Validation Accuracy Graph of Experiment 3.

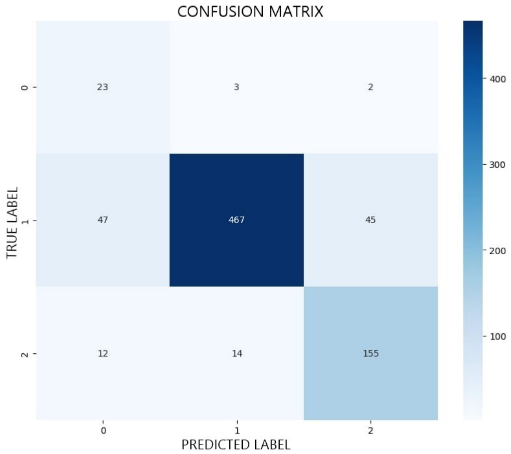

Figure 7: Confusion Matrix of a Classification Model of Experiment 3.

The orange curve (val acc) illustrates the accuracy of the validation information: It starts lower, around 22% and evolves extremely quickly between periods 0 and 1. Then, it rises around 72% and remains at this point.

The confusion matrix illustrates the results in Figure 7 of predictions made by a classifying model on a test database. The values indicate the number of elements anticipated for each category.

In negative class, 23 cases are correctly classified, while 3 and 2 are misclassified as neutral and positive, respectively.

The first class, which is neutral, is the best anticipated, and the correct predictions are illustrated by the main diagonal (23, 467, 155).

Despite the persistence of confusion between classes, it appears slightly decreased compared to the previous matrix, suggesting an optimization of the model performance, especially for classes neutral and positive.

6.4 Expriments 4 of ATFSC

In Experiment 4, The table 5 shows an accuracy of 81.36% on the CMU-MOSI dataset, achieved with an attention mechanism that enhanced information fusion. This demonstrates the model's effectiveness in emotion analysis.

Table 5: Performance of Our Model.

Model	Dataset	Accuracy	Fusion
Our model	CMU-MOSI	81.36%	Attention mechanism

Figure 8 shows the evolution of accuracy during training. The blue curve represents training accuracy, and the orange curve represents validation accuracy. The training accuracy reaches around 0.85, while the validation accuracy peaks at around 0.81, indicating a

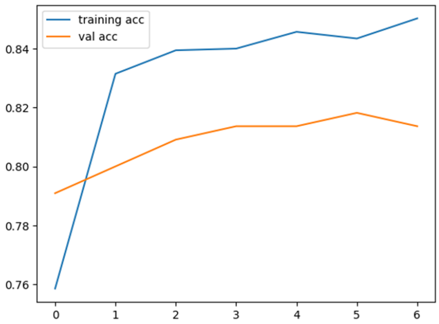

Figure 8: Training and Validation Accuracy Graph of Experiment 4.

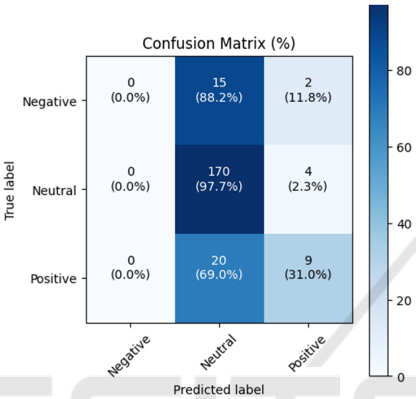

Figure 9: Confusion Matrix of a Classification Model of Experiment 4.

gap between the two curves, which limits the model's generalization ability.

As shown in Figure 9, the confusion matrix indicates a model bias towards predicting the Neutral class. 88.2% of negative, 97.7% of neutral, and 69.0% of positive samples are classified as neutral. Performance is low for the Positive class, with only 31% correctly predicted, and no predictions are made for the Negative category.

6.5 Analysis Results

Based on literature, the majority of works detect emotion recognition from text and audio modalities. But only a small number of publications highlight the need to recognize sentiments which is the interest of our paper.

For our research, we confronted the results of various sentiment identification systems from different perspectives. The multi-head approach based on transformer attention suggested by (Deng et al., 2024) achieved an accuracy of 60.74%.

(Dvoynikova and Karpov, 2023) employed a combination of Emotion HuBERT and RoBERTa to achieve an accuracy of 63.5%.

In our experimentation, we observed progressive improvements in performance across datasets. In Experiment 1, our approach, incorporating the BERT model for text and a CNN model for audio, achieved an accuracy of 64.61% on the IEMOCAP dataset. Experiment 2 demonstrated enhanced performance with an accuracy of 69%. In Experiment 3, the integration of BERT (Text) and CNN (Audio) further improved the results, achieving an accuracy of 72%. Finally, in Experiment 4, the model reached its peak performance with an accuracy of 81.36% on the CMU-MOSI dataset, leveraging an attention mechanism to enhance information fusion.

Table 6: Comparison of our approach with other works.

Works	***Model***	***Dataset***	***Accuracy***
(Deng et al., 2024)	Multi-headed attention (Transformer)	MELD, CMU-MOSEI	60.74%
(Dvoynikova and Karpov, 2023)	Emotion HuBERT + RoBERTa	CMU-MOSEI	63.5%
Our Work 2025	Bert (Text) CNN (Audio)	IEMOCAP SLUE MELD CMU-MOSI	**64.61%** **69%** **72%** **81.36%**

To the best of our knowledge, our work outperforms previous methods in terms of accuracy. These results highlight the robustness and performance of our **ATFSC** system in bimodal sentiment recognition.

7 CONCLUSION AND FUTURE WORKS

According to the literature, using single modalities does not effectively identify emotions or sentiments. This study developed a bimodal sentiment recognition system combining audio and text features, using BERT for text and CNN for audio analysis. Sentiments were categorized into negative, neutral, and positive across datasets IEMOCAP, CMU-MOSI, SLUE and MELD. An attention mechanism facilitated bimodal fusion, improving model performance from 64.61% to 81.36%.

Future work includes extending the model to recognize broader emotions and incorporating video for enhanced multimodal analysis.

REFERENCES

Bhaskar, J., Sruthi, K., and Nedungadi, P. (2015). Hybrid approach for emotion classification of audio conversation based on text and speech mining. *Procedia Comput. Sci.*, 46(C):635–643.

Deng, L., Liu, B., and Li, Z. (2024). Multimodal sentiment analysis based on a cross-modalmultihead attention mechanism. *Computers, Materials & Continua*, 78(1).

Dvoynikova, A. and Karpov, A. (2023). Bimodal sentiment and emotion classification with multi-head attention fusion of acoustic and linguistic information. In *Proceedings of the International Conference “Dialogue*, volume 2023.

Firdaus, M., Chauhan, H., Ekbal, A., and Bhattacharyya, P. (2020). Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4441–4453.

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., and Mridha, M. (2024). Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, page 100059.

Khediri, N., Ammar, M. B., and Kherallah, M. (2017). Towards an online emotional recognition system for intelligent tutoring environment. In *ACIT’2017 The International Arab Conference on Information Technology Yassmine Hammamet*, pages 22–24.

Khediri, N., Ben Ammar, M., and Kherallah, M. (2022). A new deep learning fusion approach for emotion recognition based on face and text. In *International Conference on Computational Collective Intelligence*, pages 75–81. Springer.

Khediri, N., Ben Ammar, M., and Kherallah, M. (2024). A real-time multimodal intelligent tutoring emotion recognition system (miters). *Multimedia Tools and Applications*, 83(19):57759–57783.

Lee, J. and Toutanova, K. (2018). Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 3(8).

Poria, S., Cambria, E., Howard, N., Huang, G.-B., and Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):420.

Sebastian, J., Pierucci, P., et al. (2019). Fusion techniques for utterance-level emotion recognition combining speech and transcripts. In *Interspeech*, pages 51–55.

Shon, S., Pasad, A., Wu, F., Brusco, P., Artzi, Y., Livescu, K., and Han, K. J. (2022). Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Voloshina, T. and Makhnytkina, O. (2023). Multimodal emotion recognition and sentiment analysis using masked attention and multimodal interaction. In *2023 33rd Conference of Open Innovations Association (FRUCT)*, pages 309–317. IEEE.

Wu, Z., Gong, Z., Koo, J., and Hirschberg, J. (2024). Multimodal multi-loss fusion network for sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3588–3602.

Ye, W. and Fan, X. (2014). Bimodal emotion recognition from speech and text. *International Journal of Advanced Computer Science and Applications*, 5(2).

Yoon, S., Byun, S., and Jung, K. (2018). Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE.