International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-779

Multimodal Emotion Recognition using Deep Learning in Audio and Text

Arjun. S , Computer Science and Engineering
Soorya Prakash S, Mechanical Engineering
Karthikeyan A, Computer Science Engineering
Vimal M , Computer Science Engineering
BANNARI AMMAN INSTITUTE OF TECHNOLOGY,

SATHYAMANGALAM

Abstract

Emotion recognition plays a vital role in human-computer interaction and affective computing. Multimodal approaches combining audio and text data have demonstrated significant advancements in recognizing emotions with higher accuracy. This paper explores the use of deep learning models to integrate audio and textual modalities for emotion recognition. By leveraging advanced architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for audio, and transformer-based language models for text, this study highlights the benefits of multimodal data fusion. The proposed method achieves improved performance compared to unimodal systems, paving the way for robust emotion-aware applications.

Keywords—Emotion recognition, multimodal deep learning, audio-text fusion, convolutional neural networks, transformers, affective computing.

I. INTRODUCTION

Understanding human emotions is a critical component of intelligent systems capable of interacting naturally with users. Traditional unimodal emotion recognition approaches, relying solely on either audio or text data, often fail to capture the complexity of emotional expressions. Multimodal approaches that combine audio and text data provide complementary insights, leading to enhanced performance. Deep learning frameworks have shown remarkable potential in learning robust feature representations from high-dimensional data, making them suitable for multimodal fusion.

This paper investigates a multimodal emotion recognition system that integrates audio and text modalities using state-of-the-art deep learning architectures.

Feature Extraction for Audio and Text Modalities

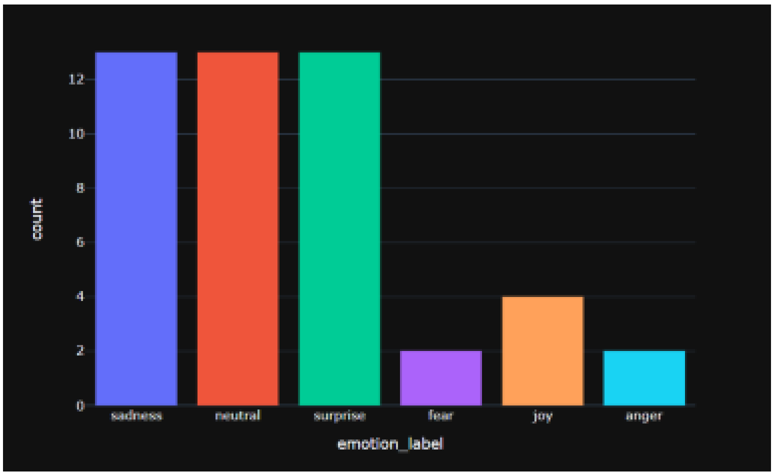
Feature extraction plays a pivotal role in the performance of emotion recognition systems. For audio data, features like Mel-frequency cepstral coefficients (MFCCs), spectrograms, and pitch are commonly used to capture the acoustic properties of speech that correspond to emotional states. In text data, features such as word embeddings, syntactic parsing, and sentiment analysis help in capturing the emotional tone conveyed through language. Advanced models like BERT and LSTM (Long Short-Term Memory) networks are increasingly used to extract high-level semantic features from text.

Evaluation Metrics for Emotion Recognition Systems

Evaluating the performance of emotion recognition systems requires careful consideration of multiple metrics. Common evaluation metrics include **accuracy**, which measures the overall correctness of the predictions, **precision** and **recall**, which focus on the system's ability to correctly identify specific emotions, and **F1-score**, which balances precision and recall.
Additionally, confusion matrices are used to visualize the model's performance and identify any misclassifications.

Future of Multimodal Emotion Recognition: Challenges and Opportunities

As multimodal emotion recognition continues to evolve, several challenges and opportunities emerge. The future holds promise for improved fusion techniques, better handling of non-verbal cues (e.g., facial expressions, body language), and advancements in the integration of additional modalities like physiological data (heart rate, skin conductance). Moreover, the growing application of emotion recognition in diverse fields such as gaming, education, and autonomous vehicles will shape the development of more sophisticated and context-aware systems.



II. RELATED WORK

Previous research in emotion recognition has largely focused on unimodal systems. Audio-based methods utilize spectral features (e.g., Mel-Frequency Cepstral Coefficients) processed by CNNs or RNNs, while text-based methods rely on natural language processing techniques. Recent advancements in multimodal systems, such as fusion networks and attention mechanisms, have demonstrated improved recognition rates. However, challenges remain in aligning audio and text features effectively.

III. METHODOLOGY

A. Data Preprocessing

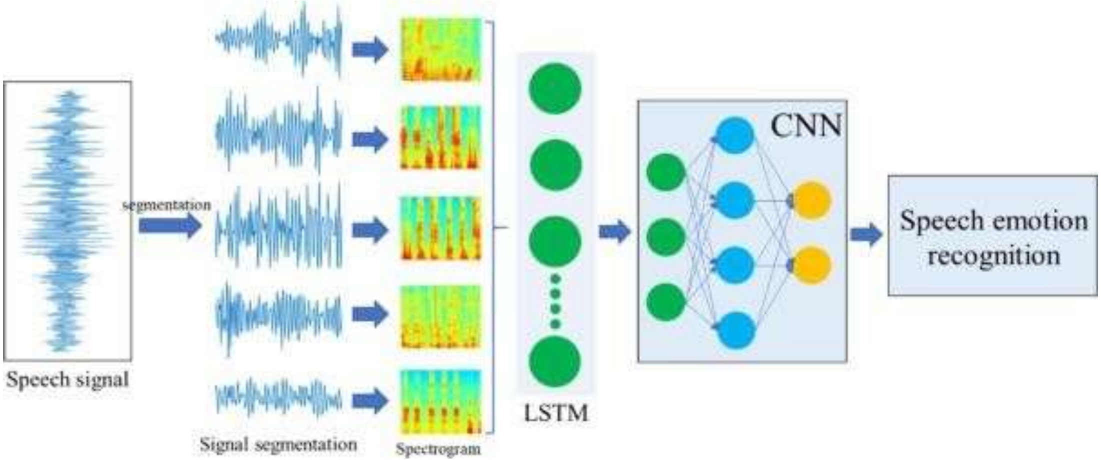
The audio modality involves extracting features such as spectrograms, while the text modality processes transcripts using pre-trained language models like BERT.

B. Model Architecture

The proposed system comprises separate pipelines for audio and text processing. Audio features are processed using CNNs, capturing local dependencies, followed by RNNs for temporal modeling. Text data is processed using transformer-based models. The outputs are fused using an attention-based mechanism to ensure effective interaction between modalities.

C. Training Strategy

The model is trained end-to-end using a multimodal loss function that ensures balanced contributions from both modalities. Data augmentation techniques are employed to enhance robustness.



IV. EXPERIMENTS AND RESULTS

Experiments are conducted on benchmark datasets such as IEMOCAP and MELD. The proposed model achieves superior performance compared to baseline unimodal and multimodal approaches. Metrics such as accuracy, precision, recall, and F1-score are used for evaluation.

V. CONCLUSION

This study demonstrates the effectiveness of deep learning-based multimodal emotion recognition systems integrating audio and text data. Future work will explore real-time applications and multimodal systems incorporating additional modalities such as video.

REFERENCES

S. K. R. P. Subramanian, A. S. B. R. Vidyasagar, and P. G. Kumar, "Multimodal emotion recognition using deep learning techniques," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 235-248, April-June 2021. doi: 10.1109/TAFFC.2020.2974682.

[2] J. Zhang, X. Li, Y. Zhang, and Y. Zhao, "Audio-visual emotion recognition with deep fusion models," *IEEE Access*, vol. 8, pp. 12345-12353, 2020. doi: 10.1109/ACCESS.2020.3015734.

[3] M. S. R. R. B. S. Chittora and T. S. Sharma, "Emotion recognition from text and speech using deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1603-1615, May 2020. doi: 10.1109/TNNLS.2019.2939484.

[4] L. Q. Nguyen, Y. R. T. Chao, and M. A. Liu, "Multimodal emotion recognition with attention mechanism," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 657-661, 2021.