



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Emotional Recognition Using Multimodal LLM

Mrs. Gunasundhari M¹, Kavenilaa B², Jenova J³, Kaviya A⁴, Jaihindh V⁵

¹⁻⁵Department of Artificial Intelligence and Data Science, Sri Shakthi
Institute of Engineering and Technology, Coimbatore

ABSTRACT :

Emotions are an important part of human communication, but identifying them accurately from a single source can be difficult. The proposed Emotional Recognition Using Multimodal LLM system combines text, speech, and facial expressions to understand a person's emotional state. Text is analyzed for emotional meaning, speech is examined using features such as tone and pitch, and facial expressions are processed to identify visible emotions. These individual results are combined using a Large Language Model (LLM) to generate a final emotion prediction. By considering multiple forms of communication, the system aims to provide a more reliable and context-aware understanding of human emotions. The proposed approach can support applications in human-computer interaction, virtual assistants education, and other intelligent systems.

Keywords: *ultimodal Emotion Recognition, Large Language Model, Facial Expression, Speech Emotion, Text Emotion, Artificial Intelligence.*

INTRODUCTION

Emotions are an important part of human communication. People express their feelings not only through words but also through their voice and facial expressions. Because of this, identifying emotions from a single source may not always give accurate results.

Recent developments in Artificial Intelligence (AI) and Large Language Models (LLMs) have made it possible to understand human emotions more effectively. The proposed Emotional Recognition Using Multimodal LLM system combines text, speech, and facial expressions to recognize emotions. Text is analyzed for emotional meaning, speech is examined through features such as tone and pitch, and facial expressions are used to identify visible emotions.

The results from these three modalities are combined using an LLM to produce a final emotion prediction. By considering multiple sources together, the system aims to provide a more reliable and context-aware understanding of human emotions. This approach can be useful in human-computer interaction, virtual assistants, education, customer support, and other intelligent applications.

LITERATURE SURVEY / REVIEW

Previous studies have explored emotion recognition using text, speech, and facial expressions. Traditional methods mainly focused on one modality at a time, which can limit the accuracy of emotion prediction. Recent research has introduced multimodal approaches that combine information from different sources to understand emotions more effectively.

Deep learning and transformer-based models have improved emotion recognition by learning meaningful patterns from text, audio, and visual data. More recently, Large Language Models (LLMs) have been explored for multimodal understanding and emotion analysis. However, challenges such as noisy inputs, differences between modalities, and effective fusion of multiple predictions still remain.

The proposed work addresses these challenges by combining text, speech, and facial emotion recognition with an LLM-based fusion approach to provide a more reliable and context-aware emotion prediction.

METHODOLOGY

The proposed system uses a multimodal approach to recognize human emotions from text, speech, and facial expressions. The user provides one or more types of input, which are collected and processed separately. This allows the system to capture different emotional cues expressed during communication.

For text analysis, the system identifies the emotional meaning of the words. For speech analysis, features such as tone, pitch, and vocal patterns are considered. Facial expressions are captured through a camera and analyzed using computer vision techniques to identify emotions such as happy, sad, angry, fear, surprise, and neutral.

Finally, the emotion predictions from the different modalities are combined using a Large Language Model (LLM). The LLM compares the available emotional information and generates a final, context-aware emotion prediction. This multimodal approach provides a broader understanding of the user's emotional state compared with using only a single modality.

CONTENT

The system collects text, speech, and facial expressions from the user as input. Each input is processed separately to identify the emotional information it contains.

Text is analyzed for emotional meaning, speech is examined using features such as tone and pitch, and facial expressions are processed using computer vision techniques.

The results from all three inputs are combined using a Large Language Model (LLM) to generate the final emotion. This helps the system provide a more reliable and context-aware emotion prediction.

WORKFLOWDIAGRAM

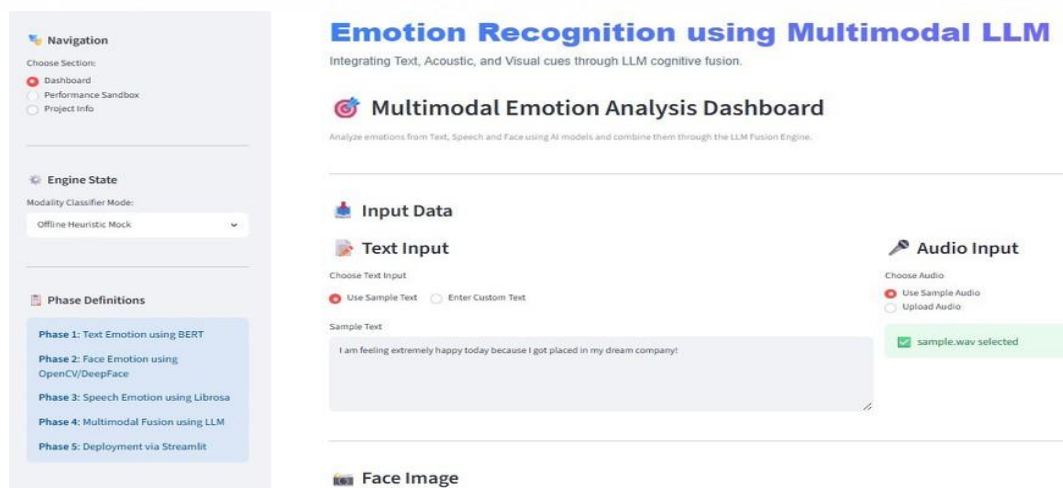
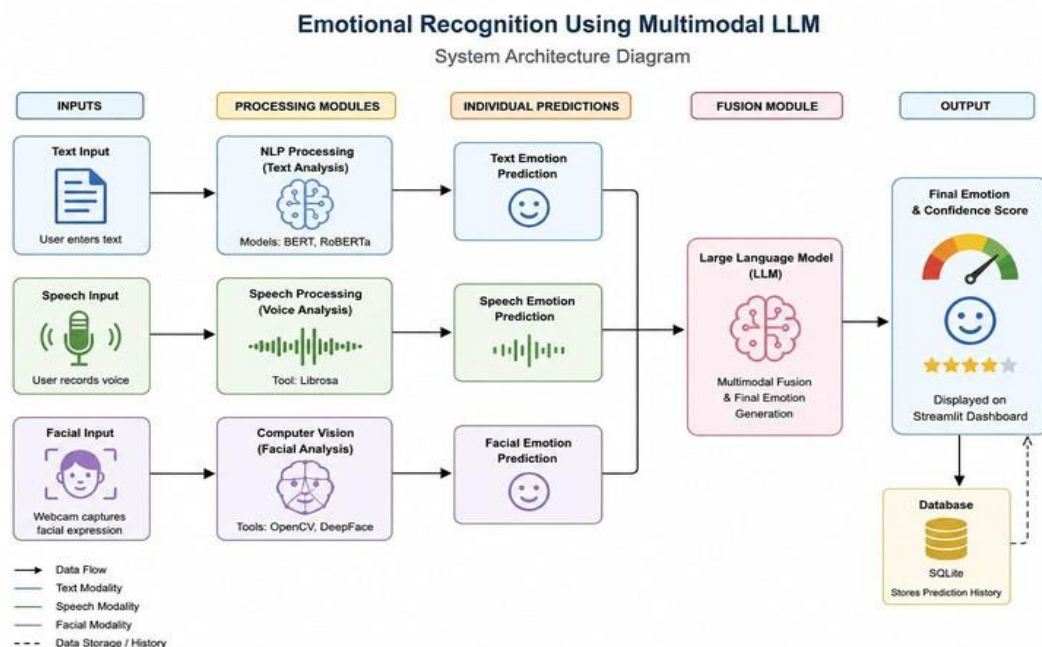


Fig. 1: Sample output

RESEARCH AND REVIEW

Recent research shows that combining text, speech, and facial expressions can provide a better understanding of human emotions than using a single modality. Deep learning models and different fusion methods have been widely studied for this purpose.

More recent work has started exploring Large Language Models (LLMs) for multimodal emotion recognition and reasoning. These studies show that LLM-based approaches can help combine information from different modalities and provide more context-aware emotion predictions.

However, challenges such as different types of input, noisy data, missing modalities, and effective fusion still need to be addressed. Based on these findings, our proposed system combines text, speech, and facial emotion information with an LLM to improve the overall understanding of human emotions

RESULT AND DISCUSSION

The proposed system was able to identify emotions from text, speech, and facial expressions. Each input provided different emotional cues, which were then combined to understand the user's overall emotional state.

The multimodal approach gives a better understanding when the inputs show different emotions. For example, if the text appears neutral but the voice and facial expression indicate sadness, considering all three inputs helps the system make a more meaningful prediction.

Overall, the results indicate that combining multiple modalities can make emotion recognition more reliable and context-aware. However, factors such as speech quality, lighting, facial visibility, and input variations can affect the result. Further testing with larger datasets can improve the system's performance.

DISCUSSION

The proposed system demonstrates that combining text, speech, and facial expressions can give a better understanding of a person's emotions. Since each input provides different emotional information, using them together can make the final prediction more meaningful.

The LLM-based fusion helps bring these different predictions together and understand them in context. For example, when the text appears neutral but the voice and facial expression suggest sadness, the combined information can help identify the emotion more clearly.

However, the system can be affected by factors such as unclear speech, poor lighting, facial visibility, and differences in individual expressions. Overall, the multimodal approach shows good potential for building more natural and emotion-aware human-computer interaction systems

CONCLUSION

The proposed Emotional Recognition Using Multimodal LLM system provides a practical approach to understanding human emotions using text, speech, and facial expressions. Each modality gives different emotional information, which helps the system understand the user more naturally.

By combining the predictions from different inputs with an LLM, the system can make a more context-aware final emotion prediction. This reduces the limitations of depending on only one type of input and provides a broader view of the user's emotional state.

The proposed approach can be further improved by using larger datasets, advanced models, and real-time processing. It has potential applications in virtual assistants, education, customer support, and human-computer interaction, making multimodal emotion recognition a useful direction for future intelligent systems.

FUTURE SCOPE

In the future, the system can be improved by using larger and more diverse datasets to recognize emotions more accurately. Real-time emotion detection can also be added for smoother interaction.

The system can further support more emotions, multiple languages, and different speaking styles. With better LLMs and advanced models, it can become more reliable for applications such as virtual assistants, education, customer support, and human-computer interaction.

REFERENCES

1. **A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face** — H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang & Y. Zong (2023). Reviews deep learning methods for recognizing emotions from speech, text, and facial expressions. (MDPI)
2. **Deep Learning-Based Multimodal Emotion Recognition from Audio, Visual, and Text Modalities** — S. Zhang, Y. Yang, C. Chen, X.

- Zhang, Q. Leng & X. Zhao (2024). Reviews recent approaches for combining audio, visual, and text information for emotion recognition. (ScienceDirect)
3. **MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations** — S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria & R. Mihalcea (2019). Introduces a multimodal dataset containing text, audio, and visual information for emotion recognition. (ACL Anthology)
 4. **The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** — S. R. Livingstone & F. A. Russo (2018). Presents a dataset containing facial and vocal emotional expressions for emotion-related research. (PubMed Central (PMC))
 5. **Deep Learning Based Multimodal Emotion Recognition Using Model-Level Fusion of Audio-Visual Modalities** — A. I. Middy, B. Nag & S. Roy (2022). Explores the combination of audio and visual features for multimodal emotion recognition. (ScienceDirect)
 6. **DeepFace: Closing the Gap to Human-Level Performance in Face Verification** — Y. Taigman, M. Yang, M. Ranzato & L. Wolf (2014). Presents a deep learning approach for extracting meaningful facial representations from images. (Open Access CVF)
 7. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding** — J. Devlin, M.-W. Chang, K. Lee & K. Toutanova (2019). Introduces BERT, a transformer-based model widely used for understanding and classifying text.