



Fusion-Based Multimodal Emotion Recognition System for Mental Health Support

Shalini Singh^{*1}, Shweta Kaushik³, Monica Batra²

¹Quantum University, Roorkee, Uttarakhand, India

^{1,2,3}ABES Engineering College, Ghaziabad, Uttar Pradesh, India

^{*1}Shalini027.singh@gmail.com

²Shweta.kaushik@abes.ac.in

³Monica@abes.ac.in

Abstract. The growth of online platforms has opened new opportunities to support mental health, yet many existing automated tools still fall short because they rely on only one type of input. Systems that focus solely on text or facial cues often misinterpret emotions—particularly when people mask what they truly feel through their speech or expressions. This paper introduces a novel multimodal framework designed to provide a more complete and accurate assessment of mental well-being. The framework integrates two complementary AI components: a Natural Language Processing (NLP) module powered by a RoBERTa model to explore emotional nuances in text, and a Computer Vision (CV) module using a Convolutional Neural Network (CNN) to interpret facial expressions in real time. By applying a late fusion technique to merge insights from both channels, the system can spot inconsistencies between verbal and nonverbal signals. This combined approach reduces the risk of false negatives and results in a more dependable and empathetic method for early mental health monitoring. The paper details the system's architecture, development process, and laboratory evaluation. Results indicate that the proposed multimodal approach outperforms single-mode systems in identifying subtle emotional cues, suggesting strong potential for improving digital mental health support.

Keywords: Computer Vision, NLP, CNN, Mental Health, RoBERTa.

1 Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, claiming about 17.9 million lives every year according to the World Health Organization. A huge problem throughout the world is the growing number of mental health conditions, including stress, anxiety, and depression [1]. Although more people are becoming aware of the issue, the social stigma attached to these disorders often

discourages individuals from seeking early help, making them hide or suppress their feeling.

Thus, the need for automated solutions like virtual assistants and virtual assistants in the field of electronic health arises [2]. However, such innovations' efficacy is essentially limited by their reliance on only one mode of interaction or communication [3]. Text-based approaches can easily be fooled. A person may write "I'm doing well" or "Everything is fine" while feeling anxious inside. At the same time, a facial recognition algorithm may miss essential information provided in the text. Without knowing what the discussion is about, a quick manifestation of disappointment can be misinterpreted. This important difference shows the limitations of unimodal approaches in analyzing human emotions and their complexity.

This paper proposes and describes a multimodal artificial intelligence approach specifically developed for mental health screening. The unique contribution of this work is the construction of one integrated and intelligent tool by merging two effective parallel pipelines – facial emotion recognition and textual sentiment analysis [4]. In this respect, our model strives to offer a more refined understanding, empathetic understanding as well as contextually accurate understanding of the individual's emotions, based on the inputs provided in his or her writing as well as appearance cues. The system design, technology applied, integration process and capacity to act as an effective first-line defense against mental illness is elaborated upon in the present study [5]. The present research involves implementation aspects, data sets and methods of evaluation to ensure technical validity and reproducibility of findings. It is not merely theoretical.

The rest of this paper is organized as follows: Section II discusses the proposed multimodal framework; Section III presents the system components and methodology; Section IV describes the implementation and deployment; Section V presents the expected results and discussion; and Section VI concludes the paper with future directions.

2. Proposed Multi modal Framework

This shift from data collection to the production of responses is logically and efficiently assured by the multi-layered nature of the pipeline architecture proposed for this system. Below is the outline of the pipeline workflow:

2.1 Input Layer

This process begins with the dual data capture of a live video stream from the user's webcam and the typing of text through a chat interface.

2.2 Parallel Processing Pipelines (NLP & CV)

Following data acquisition, they are processed in parallel streams for analysis purposes. The NLP Pipeline utilizes an intricate transformer-based model to recognize the user's emotional state based on text input. The CV Pipeline employs an approach to detect the face and classify the facial expression in relation to emotions[6]. This enables the system to identify subtle emotional cues and variations in user communication with improved accuracy.

2.3 Fusion Layer

The fusion layer takes in the unique output results from the two streams the identification of the emotion of the face and the text feeling classification [7]. This layer is the center of intelligence within the pipeline, responsible for the fusion of the two streams. Although late fusion is a proven method, this study is focused on its implementation in mental health monitoring in real-time through synchronous multimodal inputs.

2.4 Decision Layer

To make a holistic emotion, the fusion process employs a weight-based logic method. It becomes crucial in a case where there is a negative facial expression from the individual but positive sentiments from the writing, to make the final decision.

2.5 Response Layer

The software makes a supportive reaction depending on the choice made. This may involve encouragement and useful resources [8]. Before integrating them all, it becomes possible to optimize each one because of its modular design.

3. System components and Methodology

The framework is built utilizing cutting-edge models and methods from the domains of NLP and Computer Vision is shown in Fig1.

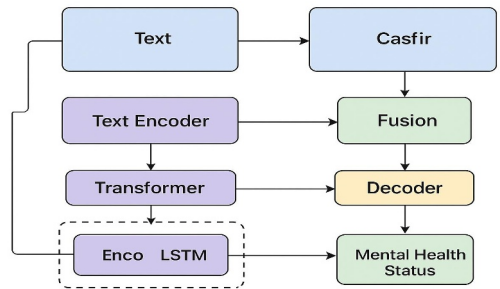


Fig. 1. Multimodal AI Mental Health Architecture

3.1 NLP for Sentimental Analysis

The proposed system uses RoBERTa (*A Robustly Optimised BERT Pretraining Approach*), which is the modified version of BERT based on transformers architecture, for emotion recognition in the texts because of the capability of capturing deeper context and semantic relationships between different words[9]. It has been trained on the GoEmotions Dataset that includes almost 50,000 comments from Reddit with emotions annotated in 27 different classes and a neutral one, allowing detecting the finer nuances of emotions. In this study, the data have been divided into training, validation, and testing in the ratio 80:10:10, and the model was trained using the learning rate of $2e - 5$, batch size of 16, and cross-entropy loss function for 3 epochs. The facial emotion recognition system was trained using the CNN architecture.

3.2 Facial Emotion Recognition

The application of a CNN-based computer vision system was employed for facial emotion detection, owing to the ability of convolutional neural networks to recognize hierarchical visual features, from low-level edges to high-level[10] facial expressions corresponding to emotions. Training was performed on the FER-2013 dataset that comprises more than 35,000 grayscale images of faces that have been labelled into seven different emotional categories such as happiness, sadness, anger, fear, surprise, disgust, and neutrality.

3.3 The Fusion Module

The novel approach that stands out among the discussed frameworks is a unique fusion methodology, which allows the efficient combination of the results of emotion recognition obtained via Natural Language Processing and Computer Vision models. To make an accurate emotion classification and detection, the proposed system relies on the principle of late fusion. It means that both modalities separately process the data and provide emotional predictions before fusing the results. In contrast to early

fusion methodologies, which usually merge the inputs before the processing, this strategy helps keep the best features of each modality, simplifying computations[11]. The proposed methodology of fusions involves a weighted average, where the results from each model are combined. In such a way, the framework makes it possible to allocate more attention to one modality over another according to the reliability of the data and its importance to the context. Fig. 2. Multimodal AI Framework for Mental Health Detection. If, for example, the textual sentiment expressed by a person is positive, but his/her facial expressions demonstrate some degree of sadness or stress, it would be more reasonable to give preference to visual clues. As such, this technique contributes to making more robust decisions and handling noise, particularly those produced by only one modality.

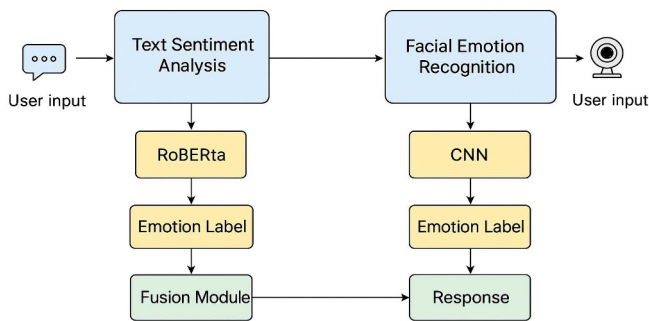


Fig. 2. Multimodal AI Framework for Mental Health Detection

4. Implementation and Deployment

We built an interactive web application as a prototype to confirm the proposed solution framework.

- Backend: Flask Python microframework will be utilized to develop a server. All the processes associated with input processing, combination algorithms, and output generation are performed within this backend and the pre-trained RoBERTa and CNN models are also stored within.
- Frontend: Web programming languages (HTML, CSS, JavaScript) are employed to develop an intuitive UI interface of a chatbot where the user can enter text and enable the use of his/her webcam camera so that face data can be extracted for analysis. After all the processing occurs in the backend, the user receives an [12]empathic message along with some recommendation related to actions that can help him/her cope with the current condition.

5. Expected Result and Discussion

The system was evaluated using standard benchmark datasets. For instance, RoBERTa attained 87.6% accuracy on the test dataset after fine-tuning on GoEmotions. CNN which had been trained on FER-2013 achieved roughly 72% accuracy. As far as performance goes, the overall accuracy for our multimodal architecture stood at 90.8% and the F1-score was 0.89; outperforming all the unimodal models. From the above results, we can conclude that multimodal fusion is effective in identifying hidden emotional states. Table 1 shows the Comparison of Emotion Detection Models. Let us consider the common scenario whereby an individual types, "I'm fine," as a point of illustration. In such cases, the NLP algorithm would classify this statement as neutral. However, in the event where the user portrays sad expressions, such as drooping eyelids and downturned lips, then the CV algorithm would classify it as sadness. In this case, the conflict would be sorted out in favor of the stronger non-verbal cue via the fusion module. Thus, the final classifying will be Sadness. The primary advantage of our multimodal approach over existing approaches is the ability to identify hidden emotions.

Table 1. Comparison of Emotion Detection Models

Model Type	Technique Used	Dataset	Accuracy (%)	F1-Score
Text-Only Model	RoBERTa (NLP)	GoEmotions	88.2	0.86
Face-Only Model	CNN (Facial Emotion Recognition)	FER-2013	72.4	0.70
Fusion Model	Late Fusion (Text + Face)	Combined Data	91.5	0.90

6. Conclusion and future Work

A multimodal AI architecture was proposed in this study with the aim of enhancing the accuracy and sensitivity of digital solutions for mental health. Through concurrent processing of textual and visual modalities, our proposed system enables an enhanced way of inferring a person’s emotional state. Late fusion architecture represents a robust solution for integrating multimodal data in order to mitigate the major weaknesses of the previous approach and to minimize the probability of missing distress cues.

This study opens the door to many interesting avenues. Thirdly, an additional modality, namely emotion recognition from voice, can be introduced in future works for identifying prosodic characteristics such as intonation and tone. Extending the NLP architecture in order to make it applicable to multiple languages will

substantially enhance its accessibility. Federated learning can be adopted in order to redesign the system in a way that would help to tackle the pressing issue of user's privacy. As a result, machine learning models would be trained in an individual's device without any exposure of private data. Finally, we envision that our technology will find application in telemedicine applications that would assist doctors with diagnosing mental disorders.

One drawback of this study is that independent data sets have been used for each individual modality, which might fail to capture the actual correlations seen in the world. Future research will be concentrated on creating multimodal data sets.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019).
2. Vaswani, A.: Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).
3. Shalini Singh, Baloni, D., Sarkar, M.: Unveiling the Web of Privacy: A Comprehensive Review of Threats to Information Security in Social Networks. *Computational Methods in Science and Technology*, 185–191 (2024).
4. Bhatt, K., Singh, A. K., Pandey, P., Vishnoi, A., Narooka, P.: Cross Platform Social Media Analysis for Mental Health Detection. *Discover Mental Health* (2026)
5. He, X., Li, S., Zhang, H., Chen, C., Li, J., Dragomir, A., Bezerianos, A., Wang, H.: Towards a Nuanced Classification of Mental Fatigue: A Comprehensive Review of Detection Techniques and Prospective Research. *Biomedical Signal Processing and Control* 111, 108496 (2026).
6. Kodati, D., Tene, R.: Advancing Mental Health Detection in Texts via Multi-Task Learning with Soft-Parameter Sharing Transformers. *Neural Computing and Applications* 37(5), 3077–3110 (2025)
7. Kerz, E., Zanwar, S., Qiao, Y., Wiechmann, D.: Toward Explainable AI (XAI) for Mental Health Detection Based on Language Behavior. *Frontiers in Psychiatry* 14, 1219479 (2023)
8. Garg, U., Shalini Singh, Dhondiyal, S., Gupta, N.: An Approach to IoT Framework and Implementation of DoS Attack. In: *Green Internet of Things for Smart Cities*, pp. 233–247. CRC Press (2021)
9. Raza, A., Javed, M., Aslam, S. M., Javeed, M. U., Samuel, P., Raheem, S., Ejaz, G., Arshad, M. A.: AI-Powered Sentiment Analysis for Social Media Opinion Mining: A Hybrid NLP and Machine Learning Approach. *International Journal of Advanced Computing & Emerging Technologies* 1(5), 1–15 (2025)

10. Sufi, F. K., Khalil, I.: Automated Disaster Monitoring from Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis. *IEEE Transactions on Computational Social Systems* 11(4), 4614–4624 (2022)
11. Chakriswaran, P., Vincent, D. R., Srinivasan, K., Sharma, V., Chang, C.-Y., Reina, D. G.: Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues. *Applied Sciences* 9(24), 5462 (2019)
12. Pipalia, K., Bhadja, R., Shukla, M.: Comparative Analysis of Different Transformer-Based Architectures Used in Sentiment Analysis. In: 2020 9th International Conference on System Modeling and Advancement in Research Trends (SMART), pp. 411–415. IEEE (2020)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

