MoodSense: A Survey on AI-Based Web Applications for Mental Well-Being and Emotional Analysis

Daivashala Deshmukh[1], Vedika Hatolkar[2], Mohini Ahale[3], Shrutika Jarwal[4]
[1,2,3,4] *Member, Maharashtra Institute of Technology Chhatrapati Sambhajinagar*

Abstract— Mental well-being is a critical component of overall health, yet many individuals face barriers in recognizing or addressing emotional distress. MoodSense is an AI-based web application designed to provide real-time emotional analysis using speech and text inputs. Leveraging Natural Language Processing (NLP) and Speech Emotion Recognition (SER), the system detects a spectrum of emotional states, including happiness, sadness, stress, and anxiety. This paper presents the architecture, methodologies, and outcomes of MoodSense, emphasizing its potential as a non-invasive, scalable, and user-centric tool for emotional awareness and mental health support. Unlike traditional tools, MoodSense offers a personalized and accessible platform suitable for students, working professionals, and general users, helping bridge gaps in mental health care through technology.

Index Terms—Mental Wellbeing, Emotion

I. INTRODUCTION

In recent years, mental health issues such as stress, anxiety, and emotional fatigue have grown significantly, particularly among students and young adults. Despite growing awareness, many individuals continue to face barriers in accessing mental health support due to stigma, cost, and limited availability of professionals. With increasing reliance on digital communication, there is an opportunity to use these modalities for early emotional assessment. MoodSense is developed to meet this need by enabling users to track and understand their emotional well-being using AI techniques that analyze natural language and vocal cues. By combining NLP for text emotion detection and SER for voice-based mood classification, MoodSense provides real-time emotional feedback in a supportive and judgment-free environment. MoodSense helps identify emotional states like stress, sadness, or happiness, and provides helpful feedback that can guide users toward feeling more balanced and in control. The goal is: to make emotional awareness more accessible. Not everyone has someone to talk to all the time. With tools like MoodSense, we hope to fill that gap, even if just a little, by using AI to support mental well-being in a way that feels human, personal, and thoughtful. MoodSense aims to bridge this gap by using AI to support emotional awareness in an intuitive, non-invasive way.

Proposed Problem Definition:

Many individuals face emotional challenges in their daily lives but often lack access to timely, effective, and personalized mental health support. Barriers such as social stigma, limited availability of mental health professionals, and the difficulty of recognizing and expressing complex emotions contribute to this gap in care. As a result, emotional distress can go unaddressed, leading to worsening mental health conditions and reduced overall well-being. MoodSense addresses this critical need by harnessing the power of artificial intelligence to analyze text and speech patterns, enabling the detection of underlying emotional states with accuracy and sensitivity. Through advanced natural language processing and voice analysis techniques, MoodSense can identify a wide range of emotions such as stress, anxiety, sadness, and happiness from everyday communication inputs. Beyond emotion detection, MoodSense offers personalized, empathetic responses tailored to the user's current emotional state, providing support, coping strategies, and recommendations for self-care. This AI-driven interaction fosters a safe and supportive environment where users feel heard and understood without fear of judgment.

Need and Motivation:

Mental health care remains out of reach for many due to limited access to therapy or counseling services.

MoodSense addresses this gap by providing an accessible, AI-powered solution that operates using just speech and text inputs. The platform offers a private and non-judgmental environment, which is especially important given the stigma that often surrounds emotional expression. Many individuals are unaware of their emotional fluctuations until they escalate; MoodSense enables real-time detection of emotions like sadness, anger, and joy by analyzing voice tone and language patterns. Using advanced NLP and SER techniques, it helps users build data-driven self-awareness by tracking emotional trends over time. As a scalable and cost-effective tool, MoodSense has the potential to deliver widespread emotional support, reaching users far beyond the limitations of traditional mental health approaches.

Objectives of Proposed Problem Definition:

The primary objective of the MoodSense project is to develop an AI-based system capable of accurately detecting and analyzing emotional states using both speech and text inputs. By employing advanced Natural Language Processing (NLP) techniques, the system identifies emotions such as happiness, sadness, anger, fear, and calmness from written content. Additionally, it incorporates Speech Emotion Recognition (SER) models that analyze vocal features such as tone, pitch, and speech patterns to classify emotions effectively. The system is designed to be non-invasive and user-friendly, ensuring that it promotes emotional awareness without relying on personal or sensitive visual data. One of the key goals is to provide real-time emotional feedback, encouraging users to reflect on and manage their mental state more consciously. Furthermore, MoodSense aims to support overall mental well-being by offering basic suggestions, mood tracking, and gentle alerts based on the emotional trends it detects over time.

Scope and Limitations:

MoodSense is primarily designed to support the mental and emotional well-being of students in higher education settings. It leverages Speech Emotion Recognition (SER) techniques to identify emotional states such as stress, anxiety, sadness, happiness, and calmness. The system facilitates real-time or scheduled check-ins by allowing students to provide short voice recordings for mood analysis. Developed as a web or mobile-based application, MoodSense can seamlessly integrate with existing student wellness platforms and assist counselors, mentors, and health professionals in recognizing students who may require additional emotional support. However, certain limitations exist. The accuracy of mood detection may be affected by factors such as diverse accents, speech impairments, background noise, or low-quality microphones. Additionally, the collection and storage of voice data raise ethical and privacy concerns, particularly regarding user consent. Furthermore, the model's performance may be biased due to training on datasets containing acted rather than real-life emotional expressions.

II. LITREATURE REVIEW

Pepino et al. (2021) Wav2vec 2.0 embeddings were used for emotion recognition in speech. The study showed strong performance without requiring large annotated datasets. The model captured nuanced acoustic features effectively. Self-supervised learning proved valuable in emotional speech modeling. Wang et al. (2024) BLSP-Emo introduced a large empathetic speech-language model. It integrates both speech and text for emotion understanding. The framework supports emotional alignment and empathy generation. It aims to enhance human-like emotional responses in AI systems. Togootogtokh & Klasen (2021) DeepEMO applies deep learning to speech emotion recognition tasks. They explored convolutional and recurrent neural networks for modeling. The system achieved good generalization on multiple datasets. Data preprocessing and architecture tuning were key to performance. Tripathi et al. (2019) A deep learning model combined audio features and transcriptions. The study leveraged MFCCs, pitch, and text sentiment features. Fusion of modalities improved recognition accuracy. Their results highlighted speech-text synergy in emotion analysis. Ekman & Friesen (1971) The study demonstrated universal facial expressions across cultures. Basic emotions like happiness and anger were consistently recognized. It laid the foundation for facial emotion research. Cross-cultural emotion perception was found to be biologically rooted. Pantic

& Rothkrantz (2003) They proposed a multimodal affect-sensitive HCI system. The framework integrated facial expressions, speech, and gestures. It aimed to improve interaction by recognizing user emotions. Early work on emotional computing interfaces was emphasized. Busso et al. (2008) Introduced the IEMOCAP emotional speech dataset. It features dyadic motion capture with emotional annotations. Designed for training and evaluating affective models. Became a benchmark in speech emotion recognition research. Schuller et al. (2013) INTERSPEECH 2013 challenged computational models on complex signals. Tasks included detecting conflict, emotion, and autism cues. It fostered innovation in paralinguistic speech processing. The challenge emphasized real-world emotional speech data. Kim & Provost (2013) Proposed pattern-based analysis of utterance-level emotion. The focus was on temporal dynamics over static snapshots. Their method captured expressive variability across speech. It improved classification of subtle emotional cues. Le & Mower Provost (2013) They used HMMs combined with Deep Belief Networks. Targeted spontaneous speech emotion recognition. The hybrid model captured both sequence and feature depth. Performance improved over traditional classifiers. Zhang et al. (2015) Introduced cooperative learning for emotion recognition from speech. Multiple classifiers collaborated to refine predictions. They emphasized knowledge sharing between models. Results showed improved robustness across datasets. Soleymani et al. (2012) Studied multimodal emotion recognition from videos. They combined audio, visual, and physiological signals. Fusion methods improved accuracy over single-modality approaches. Results supported emotion-aware multimedia applications. Alam et al. (2014) Developed laughter detection in dyadic interactions. Used multimodal features to recognize spontaneous laughter. Laughter served as a marker of social-emotional dynamics. Contributed to naturalistic emotion analysis systems. Wöllmer et al. (2008) Proposed continuous emotion recognition beyond discrete classes. They modeled long-range dependencies in emotion trajectories. The system abandoned rigid class boundaries for fluid expression. It aimed at more natural emotion representation. Yang et al. (2017) Used a double-channel neural network for facial expression recognition. Weighted feature fusion improved recognition accuracy. Focused on static and dynamic facial cues. The method addressed both intensity and variability in expressions. Zadeh et al. (2016) Presented the MOSI multimodal opinion video corpus. It includes sentiment and subjectivity annotations. MOSI supports fusion of text, audio, and video features. A valuable resource for affective content analysis. Cohn et al. (2007) Explained facial expresson coding via the FACS system. Human annotators rated expressions based on action units. FACS became a foundation for automatic facial analysis. Widely used in psychological and computational studies. Lucey et al. (2010) Released the CK+ dataset for facial expression research. Included posed expressions with action unit labels. Supported development of expression recognition models. Became a standard for training and validation. Gunes & Pantic (2010) Surveyed dimensional emotion modeling in recognition systems. Focused on continuous prediction of valence and arousal. The paper highlighted the limits of discrete emotion labels. Promoted a more realistic representation of emotional states. Eyben et al. (2010) Developed openSMILE, an audio feature extraction toolkit. Supports real-time and batch feature computation. Widely adopted in affective computing research. Enables reproducibility and scalability in signal analysis. Ringeval et al. (2013) Introduced the RECOLA corpus for remote affective interaction. Captured audio, video, and physiological signals.. Schuller et al. (2012) Launched the AVEC challenge on continuous emotion prediction. Participants modeled emotions from audiovisual signals. Baltrušaitis et al. (2016) Released OpenFace, a toolkit for facial behavior analysis. Tracks facial landmarks and estimates emotions. Supports both academic and applied research use. Open-source availability promotes reproducibility. Devlin et al. (2019) Introduced BERT, a transformer-based language model. Used bidirectional context for understanding semantics. Achieved state-of-the-art across multiple NLP tasks. A foundation for emotion recognition in text. Radford et al. (2019) OpenAI's GPT models used unsupervised multitask learning. Demonstrated generalization without task-specific training. Language models learned rich representations from large data. Enabled new directions in emotional and contextual AI.

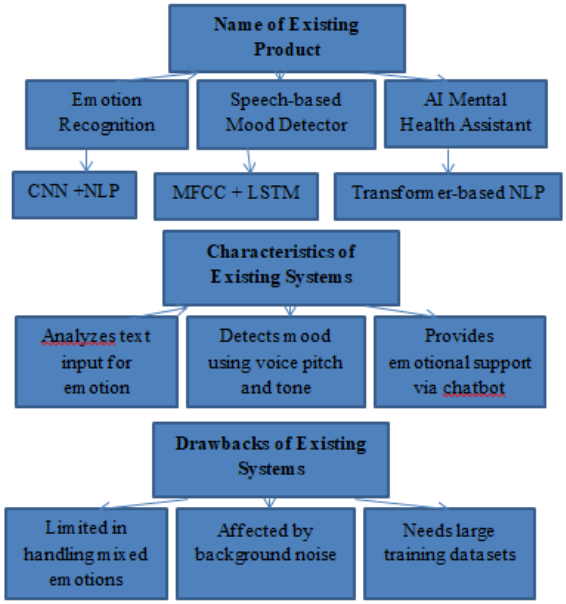

Fig 1: Literature Review based on Existing Product

III. PROPOSED RESEARCH METHODOLOGY

The methodology includes four phases: input collection, preprocessing, emotion detection, and feedback generation

Input Data Collection: This step involves input data collection through text or voice, analyzing things like the words used or the way something is said.

Preprocessing Emotion Detection: This step involves cleaning up and preparing the data from the sensors or inputs

Feedback: Once the mood is detected, the system can display the results on a screen, send a notification, or even trigger a response (like playing a song or suggesting an activity based on the detected mood).

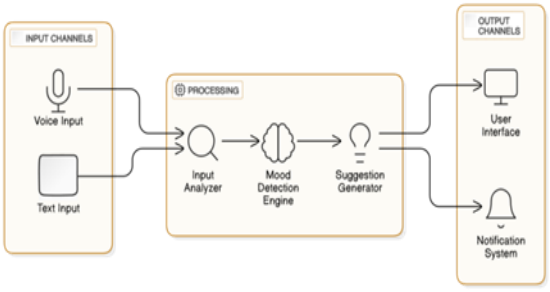

Fig 2: Operational low of MoodSense - From Input to Emotional Insight

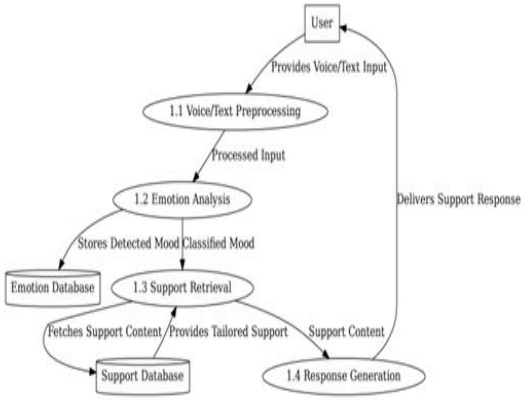

Fig 3: Architectural Overview

The MoodSense architecture is divided into four key modules:

1. User Interface Layer:

-Provides a front-end interface for users to input text or voice data. - Supports both web and mobile platforms. - Displays emotional feedback and suggestions to users in a simple, user-friendly format.

2. Input Processing Module:

- Handles both text and speech data. - Text inputs are cleaned, tokenized, and vectorized using NLP preprocessing steps. - Speech inputs are converted into spectrograms and MFCCs for analysis.

3. Emotion Detection Engine:

- Utilizes two parallel AI pipelines: - NLP-Based Emotion Classifier: Uses transformer-based models (like BERT) to analyze text for emotional tone. - Speech Emotion Recognizer: Uses deep learning (CNN/LSTM) to identify emotion from voice pitch, tone, and tempo. - Results from both models are fused to provide a holistic emotional state classification.

4. Feedback & Recommendation System:

- Translates emotion classification into meaningful suggestions. - Offers personalized well-being advice like relaxation tips, journaling prompts, or music playlists. - Stores anonymized emotion trends for long-term analysis (optional and consent-based). The system ensures privacy and ethical data handling, avoiding personal identifiers.

IV. DISCUSSION & IMPLICATION

MoodSense offers promising applications across education, workplaces, and general use cases. However, challenges such as data privacy, speech variability, and model bias require careful consideration. Future development may include expanding emotion categories, incorporating facial emotion detection, and improving real-world model generalization.

V. CONCLUSION OF FINDINGS

MoodSense demonstrates the effective use of AI in mental health support. Through real-time analysis of speech and text, it delivers accessible emotional insights and helps users track and understand their mental states. The tool bridges the gap in emotional support systems using scalable and ethical technology solutions.

REFERENCE

[1] Pepino, Leonardo, et al. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. arXiv:2104.03502, arXiv, 8 Apr. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2104.03502.

[2] Wang, Chen, et al. BLSP-Emo: Towards Empathetic Large Speech-Language Models. arXiv:2406.03872, arXiv, 6 Jun. 2024. arXiv.org, https://doi.org/10.48550/arXiv.2406.03872.

[3] Togootogtokh, Enkhtogtokh, and Christian Klasen. DeepEMO: Deep Learning for Speech Emotion Recognition. arXiv:2109.04081, arXiv, 9 Sep. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2109.04081.

[4] Tripathi, Suraj, et al. Deep Learning Based Emotion Recognition System Using Speech Features and Transcriptions. arXiv:1906.05681, arXiv, 11 Jun. 2019. arXiv.org, https://doi.org/10.48550/arXiv.1906.05681.

[5] Ekman, Paul, and Wallace V. Friesen. 'Constants across Cultures in the Face and Emotion.' Journal of Personality and Social Psychology, vol. 17, no. 2, 1971, pp. 124–29. DOI.org (Crossref), https://doi.org/10.1037/h0030377.

[6] Pantic, M., & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE, 91(9), 1370-1390.

[7] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335-359.

[8] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Moosmayr, T. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. Proceedings of INTERSPEECH, 148- 152.

[9] Kim, Y., & Provost, E. M. (2013). Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. Proceedings of ICASSP, 3677-3681.

[10] Le, D., & Mower Provost, E. (2013). Emotion recognition from spontaneous speech using Hidden Markov Models with deep belief networks. Proceedings of ICASSP, 3716-3720.

[11] Zhang, Z., Coutinho, E., Deng, J., & Schuller, B. (2015). Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1), 115-126.

[12] Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing, 3(2), 211-223.

[13] Alam, F., Riccardi, G., & Sarkar, P. (2014). Multi-modal laughter detection in naturalistic dyadic interactions. Proceedings of INTERSPEECH, 2585-2589.

[14] Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes—Towards continuous emotion recognition with modelling of long-range dependencies. Proceedings of INTERSPEECH, 597-600.

[15] Yang, B., Cao, J., Ni, R., & Zhang, Y. (2017). Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE Access, 6, 4630-4640.

[16] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online

opinion videos. IEEE Transactions on Affective Computing, 9(4), 496-508.

[17] Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. In J. A. Coan & J. J. B. Allen (Eds.), Handbook of emotion elicitation and assessment (pp. 203-221). Oxford University Press

[18] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 94-101.

[19] Gunes, H., & Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions, 1(1), 68-99.

[20] Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, 1459-1462.

[21] Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 283-288.

[22] Schuller, B., Valstar, M., Eyben, F., Cowie, R., & Pantic, M. (2012). AVEC 2012: The continuous audio/visual emotion challenge. Proceedings of the 14th ACM International Conference on Multimodal Interaction, 449-456.

[23] Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. Proceedings of IEEE Winter Conference on Applications of Computer Vision, 1-10.

[24] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171-4186.

[25] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.