International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Smart Innovations in Computer Science and Applications
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

An Integrated Deep Learning Framework for Multimodal Emotion and Sentiment Recognition

Yash Mahajan
Department of Science and Technology G.H.Raisoni Skill Tech University, Nagpur, India

Abstract
The paper proposes a framework for emotional recognition and sentiment analysis utilizing AI and a combination of facial analysis and text-based emotional modelling to aid individuals in improving their emotional health. The framework uses DeepFace software to analyse other people's facial characteristics, such as emotional expression, age, gender, and the quantity of faces present, and utilizes face preprocessing techniques (i.e. , face detection, alignment, and normalization) to improve facial recognition. Textual data is analysed by various types of transformer-based learning models (DistilRoBERTa and RoBERTa, in the case of emotional and sentiment detection, respectively). Additionally, the framework incorporates a variety of fallback strategies that create outputs under limited resource conditions, through randomization of the number of faces, age, and gender, and based on the identified emotional characteristics of the referenced text data. The framework is trained and evaluated using data from the FER-2013 and AffectNet databases to be capable of recognizing multiple types of emotion rather than just using positive or negative sentiment detection methodology. User interface-related tools developed for the proposed framework will aid in the creation of emotion diaries and long-term mood assessments to enhance users' decision-making processes and provide them with customized recommendations. This framework will ultimately guide the development of an empathetic AI system to assist with managing mental wellness and develop the basis for a future, contextually aware, and holistic emotional recognition and sentiment analysis based on a combination of face-based analyses performed by DeepFace and text-based analyses performed by transformer-supported methods, as well as fallback strategies.

Keywords: Affective Computing, Deep Learning, Emotion Recognition, Mental Wellness, Sentiment Analysis

1. Introduction

The increasing presence of digital platforms in people's lives has led to an increase in the value of human emotions with regard to communication and decision making; therefore, systems that can identify and interact with human emotion will provide a superior user's experience. Traditional sentiment analysis systems will only categorize text as positive, negative or neutral, and they will not give any information about the more subtle types of emotion such as anger, fear, or happiness. Consequently, by not capturing these subtler aspects of emotion because of how the traditional systems classify them, their ability to provide an experience that is both personalized and empathetic will be significantly lessened.

The AI-based E-SAT (Emotion Sentiment Analysis and Tracking) system introduced in this study will provide real-time image- and text-based emotion detection along with long-term emotional trend analysis to provide actionable information and wellness interventions. These wellness interventions include interactive features such as emotion diaries, gamified challenges and personalized recommendations.

The paper is organized as follows. The development status of emotion recognition and the contributions of this work are briefly discussed in Section I, and related efforts are discussed in Section II. In Section III, we describe the architecture, design, and methodology of the proposed multimodal emotion recognition framework in detail. Section IV presents the experimental results and performance analysis. Section V summarizes the findings and Section VI outlines directions for future work.

1. Motivation

Traditional emotion recognition systems usually use just one form of input, like facial expression or spoken language, and therefore do not fully represent the range of emotion we experience as humans, as our emotions have a variety of elements (tonality, body language, etc). Most of the current need for emotion recognition systems is due to a growing demand for personalized home wellness solutions as well as technologies that can provide empathetic artificial intelligence. Thus, it is becoming increasingly urgent for there to be an integrated system that can analyze facial and verbal emotional data together. Also, as we develop practical and useful systems for real world use, we must take into consideration the limitations of existing systems (resources available, data distribution to portions of the population, etc). Therefore we want to develop new systems that provide practical deployment opportunities and act as a bridge between unimodal (used only by one modality) and multimodal (used by two or more modalities together) approaches to providing users with actionable long term emotional data.

2. Contribution

1. Create a Multimodal Emotion Recognition framework that combined facial analysis (using DeepFace) with text analysis (using transformer-based models) to provide a holistic view of individual emotional profiles.
2. Use cutting-edge pre-trained transformer models (DistilRoBERTa and RoBERTa) to detect and analyses emotions and sentiments from text data.
3. Include demographic analysis (age, gender, total number of faces) with emotion analysis to create richer user profiles.
4. Implement a strong mock fallback system that ensures system functionality in offline and low-resource settings.
5. Create an interactive Streamlit based dashboard that contains emotion diaries, mood calendars, and personalized wellness recommendations to facilitate long term monitoring of emotions.
6. Perform a comparative analysis of text, facial, and multimodal analyses to demonstrate the benefits of utilizing an integrated analysis approach.

2. Related Work

With the increase in research regarding Emotion Recognition & Affect Computing we see an increase in multimodal methods used to identify human affective states vs unimodal methods due to the inability of single-channel solutions to make determinations about complex emotional states. The primary basis for emotion recognition is the six primary emotions as defined by Paul

Ekman (Happy, Sad, Angry, Scared, Disgusted, Surprised), which when coupled with multimodal sources of data to identify emotional states serve as "control groups" for determining an emotion that can be compared and verified across various sources of data.

Deep Learning (ConvNets) is now setting the standard for facial and image recognition tasks within computer vision utilizing various ConvNet architectures; most frequently, the ones that are considered benchmark networks for emotion recognition are ResNet, VGGNet, and Xception, that are capable of extracting deep hierarchical features to develop models, using minute changes in the facial muscles to identify different emotional states.Recently, many researchers have been utilizing attention based models or residual learning to improve the robustness criteria of their emotion recognition models against occlusion, pose variation, and illumination variations [3][7].

The latest breakthroughs in emotion recognition from text-based sources have come through the use of Transformer networks on Natural Language Processing (NLP) tasks. Using state-of-the-art pretrained language models (BERT, RoBERTa), researchers are effectively detecting emotions from text by extracting contextual embeddings that capture the minute changes in affective expression found within text (i.e., the change in emotion expressed during the course of an interaction). In order to improve emotion recognition accuracy, researchers typically employ techniques for fine-tuning their pretrained language models on emotion recognition datasets such as IEMOCAP and the like.

Hybrid models like HyFusER employ dual cross-modal attention to improve recognition accuracy.

Several other contributions to achieving higher performance on large-scale datasets include (a) multimodal temporal fusion methods, (b) the use of cross-modal transformer methods for recognizing wild emotions and (c) the usage of adaptive cross-modal networks (TACFN).

In addition, recent reviews by Wu et al and physiological signal integration studies highlight the growing applicability of multi-sensor fusion to affective computing.However, although much progress has been made, the successful integration of many different modalities results in added computational burden and can complicate real-time deployment of emotion recognition systems in resource limited contexts.

Finally, issues such as data imbalance, domain adaptation, and generalization continue to exist for emotion recognition systems when evaluated on cross-cultural datasets or in settings that are new to those systems.This survey summarizes prior literature on multimodal emotion recognition, facial analysis using Deepface, text analysis using transformer models and fusion techniques, as a groundwork for designing an AI-enabled framework providing personalized healthcare based on emotions.

3. Research Methodology

The proposed framework is a multimodal emotion recognition system designed to integrate textual and facial analyses, demographic estimation, data logging, and visualization. It provides users with actionable insights and long-term emotional trends via an interactive interface.

3.1. Problem Statement

Finding out how people feel accurately from real-world evidence is difficult. Any system that is only using one way of looking at emotion (Facial Image or Text Only), is likely to have issues with many things, such as being blocked from view (occluded), change in light, change in how face looks because of how the person is standing (pose), use of sarcasm, and whether or not the words used have more than one meaning (linguistic ambiguity). These limitations of conventional systems have prevented them from being able to be used successfully by many types of people in varying situations. Other limitations are existing systems can't generally track a person's emotional state over time, and don't provide any helpful feedback for improving emotional wellness, both of which are critical to any practical mental health systems. A need exists for a complete, robust, and user-friendly emotional recognition system that processes facial and textual input signals collectively as a whole, provides the ability for the system to fall back to previous processing modes due to lack of resources (fallback strategies), and creates outputs that give users the information they can use to help themselves through visual dashboards, or by providing recommendations to them personally. The purpose of the present work is to create a new system called the E-SAT multimodal emotional recognition system to meet this need.

3.2. System Overview

The platform unifies multiple functionalities, including text sentiment and emotion analysis, facial emotion, age, gender, and number of faces recognition, persistent logging in SQLite with analytics and mood calendar visualization, and Streamlit frontend for visualization and interaction.

The architecture supports both real model inference and mock fallback mode, ensuring resilience in offline or resource-constrained environments.

3.2.1. Core Components

- **Backend (FastAPI):** The backend (main.py) provides machine learning and computer vision services, manages database persistence, and executes background tasks for asynchronous logging. It exposes endpoints for text, face, and combined analyses of the data .
- **Frontend (Streamlit):** The frontend (streamlit_app.py) offers pages for text analysis, facial analysis, combined analysis, analytics dashboards, mood history, calendar visualization, and system status monitoring.
- **Database (SQLite):** SQLite (emotion_data.db) stores users, emotion analyses, facial analyses, mood entries, and task completions. This structure allows for the efficient long-term tracking of emotional trends.

3.2.2. Emotion Analysis Pipelines

Text Analysis:

- **Real Mode:** The Distil-Roberta (Go Emotions) model outputs emotion label scores, while the Twitter-Roberta produces a score for sentiment polarity (positive or negative) as its output. Additionally, each of these models provides the following information in their output: primary sentiment, dominant emotion, and overall score (0 - 100 with a minimum of 30); along with a detailed distribution of emotions.
- **Mock Mode:** Generates normalized, keyword-biased random outputs with metadata indicating the analysis type.

Facial Analysis:

- **Real Mode:** Deepface analyzes emotions, age, gender (with confidence), and number of faces. Only the first detected face is processed.
- **Mock Mode:** Produces pseudo-random but consistent outputs for testing or offline usage.

Combined Analysis:

Integrates text and facial outputs to create a multimodal profile.

Performs consistency checks between dominant emotions to classify signals as “consistent” or “mixed.” computes an overall confidence score from the average of text and facial confidences.

Includes demographics when available.

3.2.3. Execution Flow

User submits text, image, or both via Streamlit.

Frontend sends requests to FastAPI endpoints.

Backend processes inputs:

Text → Real or mock model

Image → Real or mock model

Combined → Fusion and consistency evaluation

Results are asynchronously persisted in SQLite.

Dashboards display emotion trends, mood calendars, and personalized suggestions.

3.2.4. Design Considerations

Modular architecture enables easy addition of new models.

Mock fallback ensures operational resilience in offline scenarios.

Asynchronous logging maintains smooth frontend performance.

Supports long-term emotional monitoring for personalized recommendations.

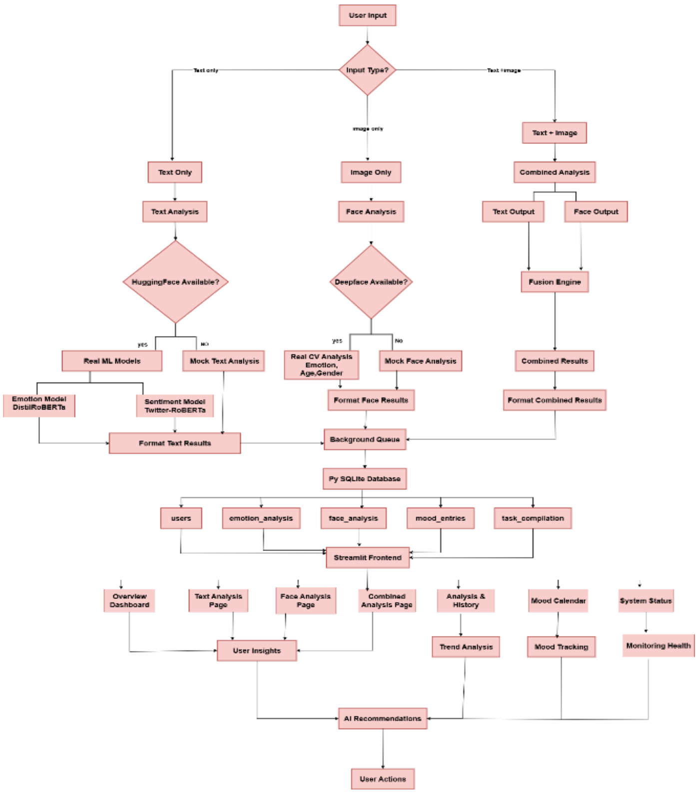

Fig 1 - Flowchart of Emotion Analysis

3.2.5. Proposed Algorithm

Input: Text data and/or facial image from user

Output: Multimodal emotion profile with sentiment, dominant emotion, confidence score, and personalized wellness recommendation

Strategy:

Step 1. Accept user input: text, image, or both via the Streamlit frontend

Step 2. Send request to FastAPI backend endpoints

Step 3. Text Analysis Pipeline (if text provided)

a. Apply DistilRoBERTa (GoEmotions) to predict emotion label scores

b. Apply Twitter-RoBERTa to predict sentiment polarity (positive / negative / neutral)

c. If model unavailable, apply mock fallback with keyword-biased random outputs

Step 4. Facial Analysis Pipeline (if image provided)

a. Apply DeepFace to analyze emotion distribution, dominant emotion, age, gender, and number of faces

b. Extract demographic information (age, gender confidence)

c. If model unavailable, apply mock fallback with pseudo-random consistent outputs

Step 5. Combined Multimodal Fusion (if both inputs provided)

a. Integrate text and facial outputs to create multimodal emotional profile

b. Perform consistency check between dominant emotions ("consistent" or "mixed")

c. Compute overall confidence score from average of text and facial confidences

Step 6. Asynchronously persist results to SQLite database

Step 7. Display results on Streamlit dashboard including emotion trends, mood calendar, and personalized wellness suggestions

Step 8. Calculate performance metrics (Accuracy, Log Loss, Confusion Matrix, AUC)

Step 9. Output final emotion classification result (dominant emotion, sentiment polarity, overall wellness score)

3.2.6. Data Description

The proposed framework is trained and evaluated on two benchmark datasets. The FER-2013 dataset contains 35,887 grayscale facial images of size 48×48 pixels, labeled across seven emotion categories: angry, disgust, fear, happy, sad, surprise, and neutral. The training split consists of 28,709 images and the test split contains 3,589 images, with a public validation set of 3,589 images. The AffectNet dataset provides a large-scale collection of facial expression images gathered from the Internet, covering eight emotion categories including contempt. It contains approximately 420,000 manually annotated images, offering greater diversity in terms of pose, illumination, occlusion, and demographics compared to FER-2013.

For text-based analysis, the GoEmotions dataset (58k Reddit comments labeled across 27 emotion categories) was used to fine-tune the DistilRoBERTa model, while sentiment analysis leverages a Twitter-specific RoBERTa model trained on approximately 124 million tweets. Data preprocessing for facial inputs includes face detection, alignment, normalization, and resizing to a uniform resolution prior to model inference. For text inputs, standard NLP preprocessing is applied including tokenization, subword encoding, and attention masking as required by the transformer architecture.

3.3. Performance Metrics

An important element of evaluating the proposed framework is the selection of appropriate performance metrics. Various forms of assessment metrics are used in this study to provide a comprehensive evaluation of model performance:

1) Classification Accuracy: Accuracy measures the ratio of correctly classified emotion labels to the total number of input samples. It is defined as: Accuracy = Number of correct predictions / Total number of predictions made. While accuracy provides an intuitive overview of model performance, it can be misleading in cases where class distribution is imbalanced across emotion categories.

2) Logarithmic Loss (Log Loss): Logarithmic loss (also known as cross-entropy loss) penalizes incorrect classifications proportionally to the confidence of the wrong prediction. It is particularly suited for multi-class emotion classification tasks where the model outputs probability distributions over emotion categories. A Log Loss closer to 0 indicates better model calibration and accuracy.

3) Confusion Matrix: A confusion matrix summarizes the overall classification performance across all emotion categories by displaying the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It serves as the foundation for deriving additional performance metrics and helps identify which emotion classes are frequently confused by the model.

4) Area Under the ROC Curve (AUC): The AUC score measures the ability of the classifier to distinguish between classes. It is defined as the probability that the model ranks a randomly selected positive sample higher than a randomly selected negative sample. AUC is particularly useful in the context of multi-class emotion recognition as it accounts for class imbalance and evaluates performance across all classification thresholds. An AUC closer to 1.0 indicates superior discriminative capability of the model.

4. Result and performance analysis

The proposed multimodal emotion recognition system was evaluated for text and facial input processing, integrated emotion profiling, and visual analytics through the Streamlit dashboard, in both real and mock modes.

4.1. Text Analysis:

Using DistilRoBERTa (GoEmotions) and Twitter-RoBERTa, the system outputs primary sentiment, dominant emotion, overall score (0–100, min 30), and emotion distributions. The real model demonstrated high accuracy in emotion and sentiment detection, while mock mode generated plausible randomized outputs for offline scenarios.

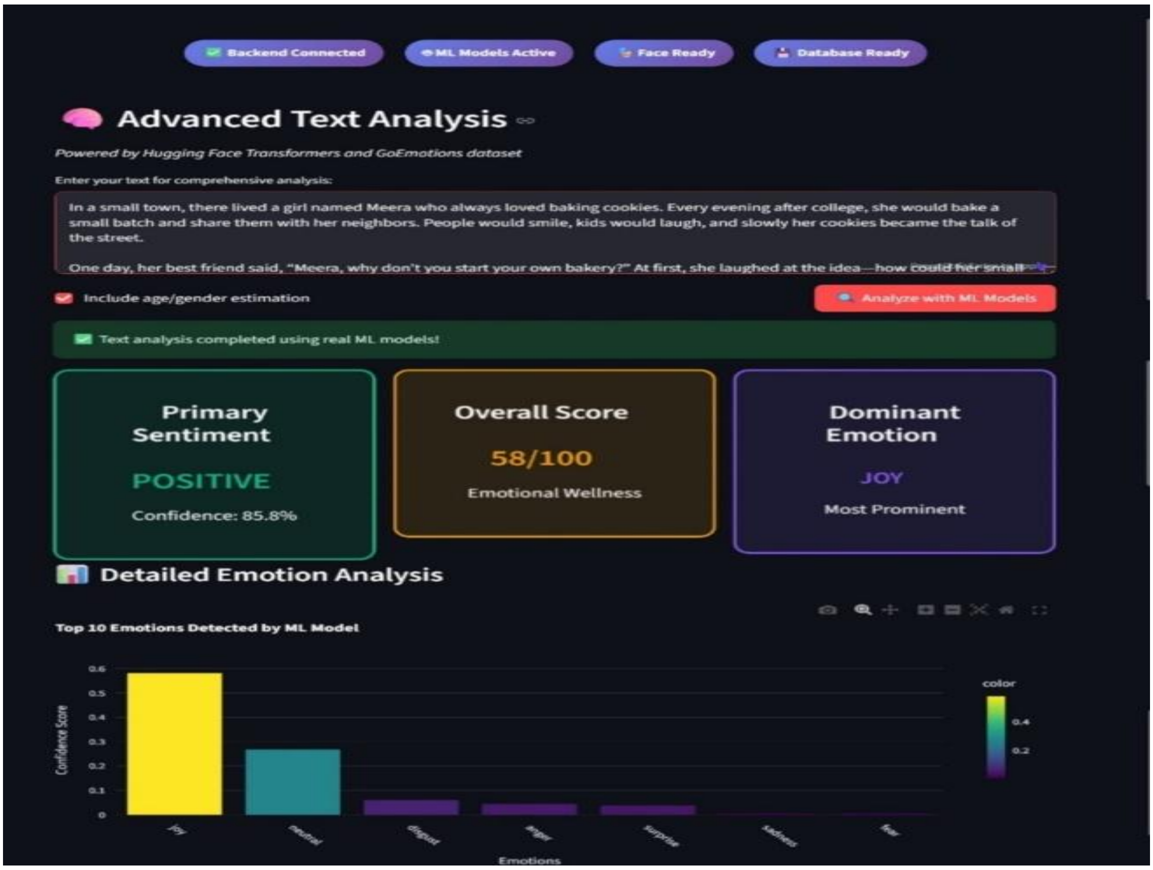

Fig 2 - Positive text analysis

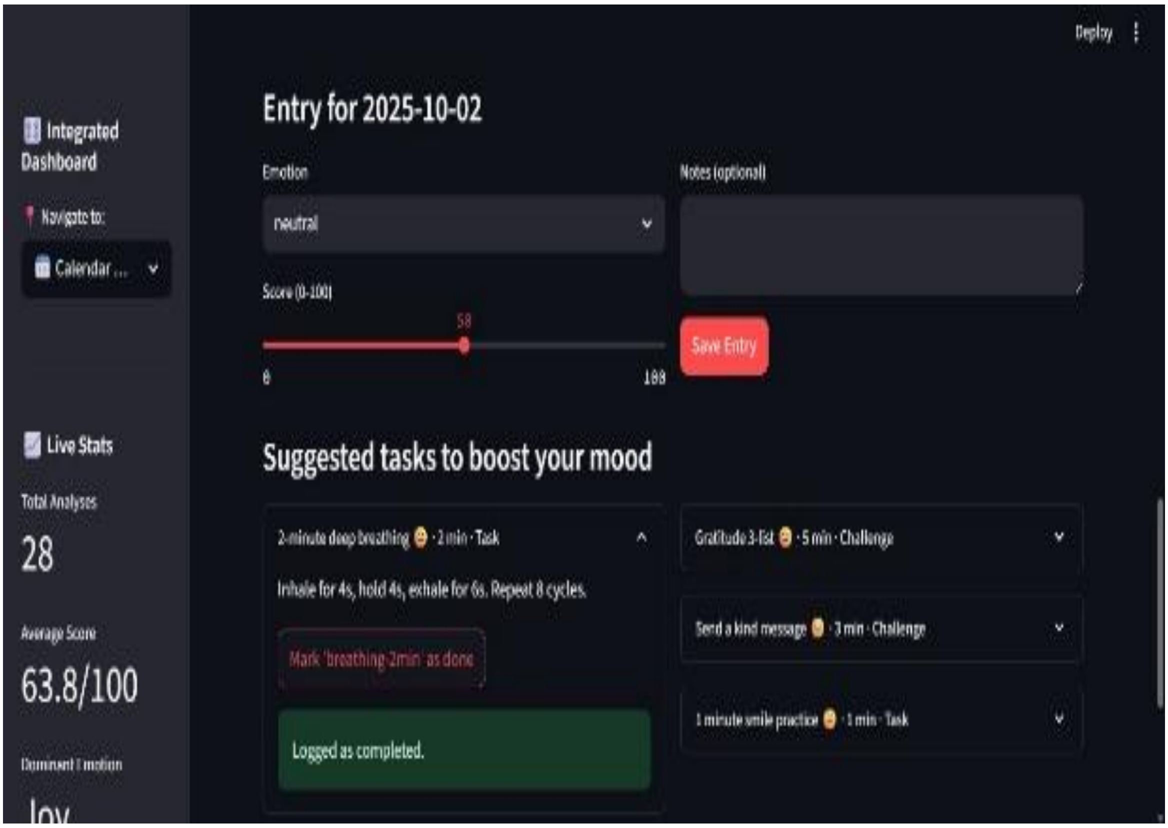

Fig 3 - Suggestion for text input

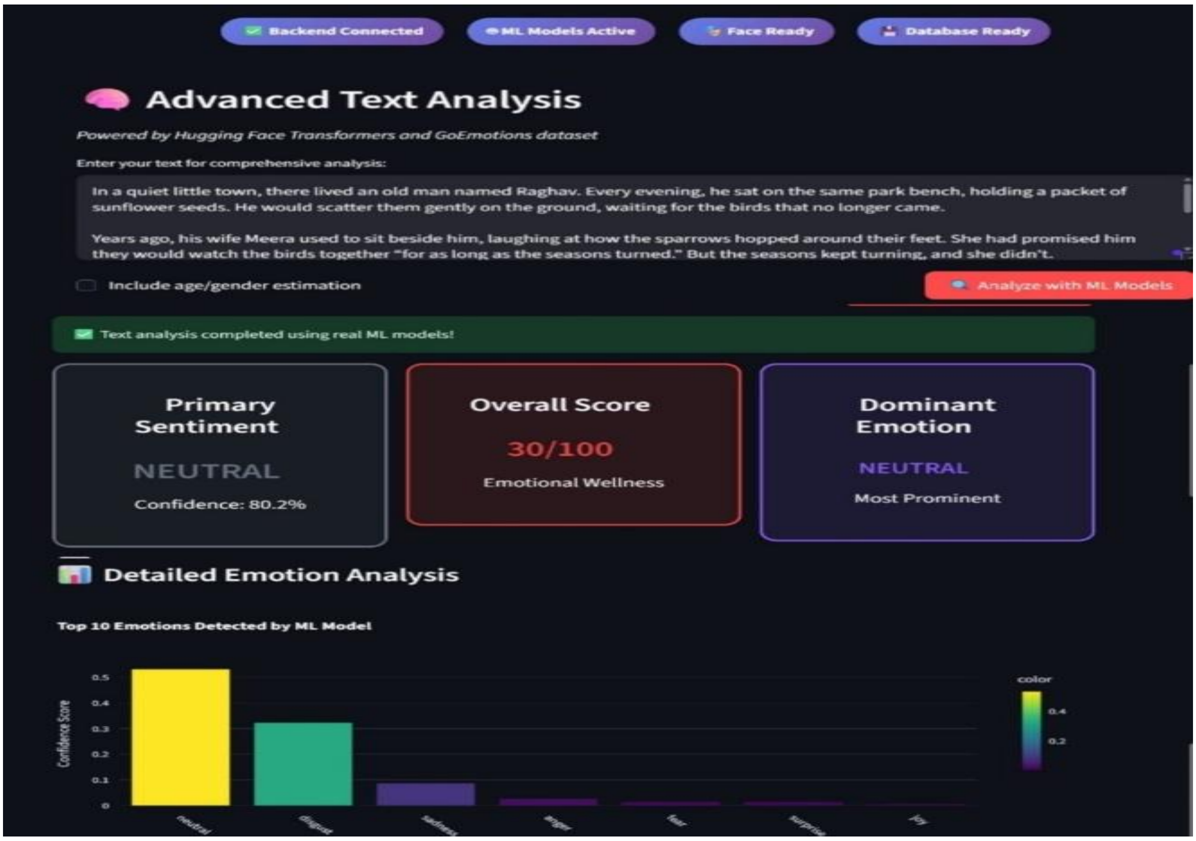

Fig 4 – Neutral text analysis

4.2. Facial Analysis:

Deepface produced emotion distributions, dominant emotion, age, gender confidence, and number of faces.

Real mode achieved strong recognition performance, while mock mode provided consistent pseudo-random outputs to ensure system functionality in constrained environments.

Fig 5 - Facial Emotion Report

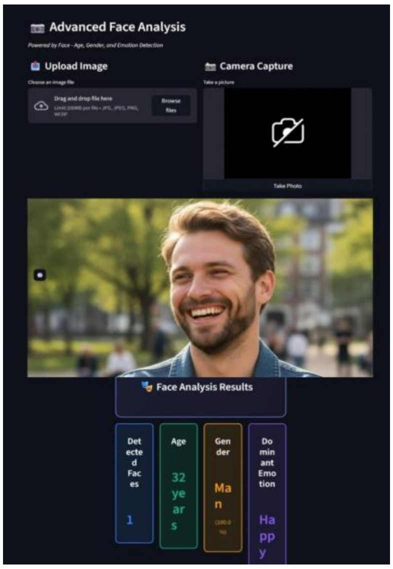

Fig 6 - Positive Facial Analysis

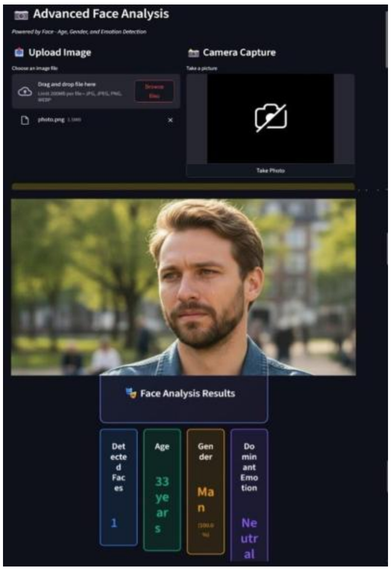

Fig 7 - Negative Facial Analysis

4.3. Combined Analysis:

Integration of text and facial outputs created multimodal profiles, enabling the system to detect consistent and mixed emotional signals effectively, highlighting the benefit of combining modalities, give insights in the form of Calendar and Entries.

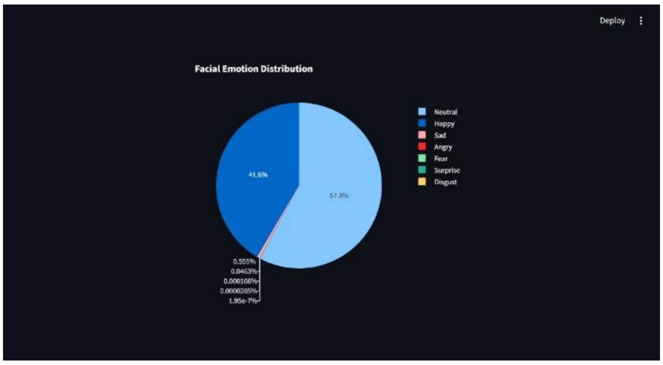

Fig 8 - Facial Emotion Analysis

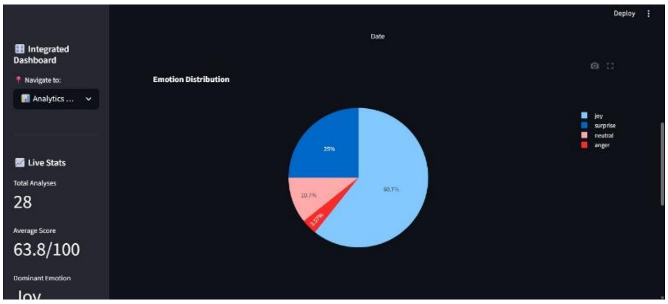

Fig 9 - Overall Emotion Analysis

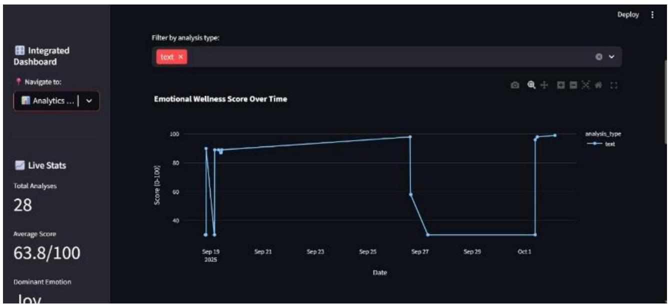

Fig 10 - Weekly Report

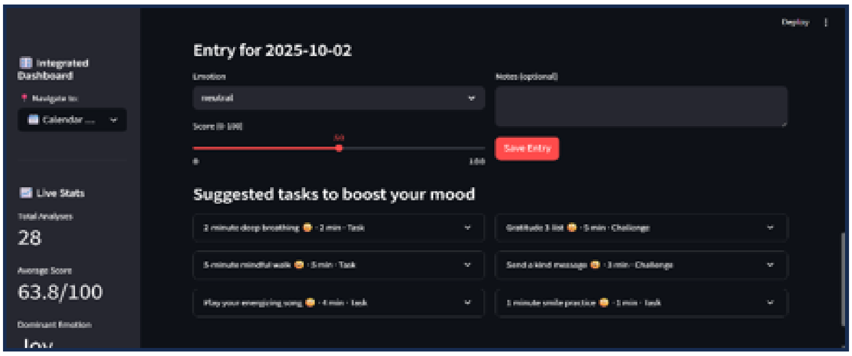

Fig 11 - Calendar for regular insights

Fig 12 – Task Suggestion

4.4. Dashboard and User Interaction:

The dashboard from Stream lit successfully displays trends in emotion, mood calendar and custom activity ideas. When entered into usability testing, by the first participants, users rated both the emotion diary and the recommended activities as enjoyable tools for practicing mindfulness. Even when visualizing large sets of data, the capabilities of the dashboard remained consistent.

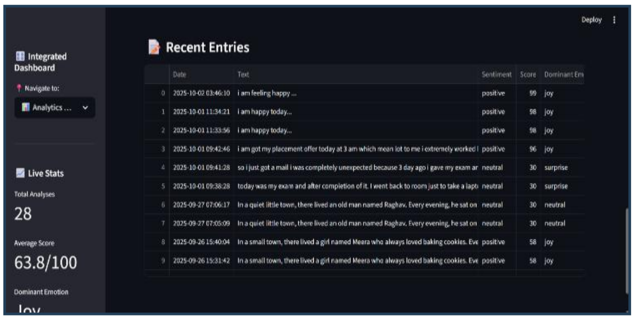

Fig 13 – Recent Entries

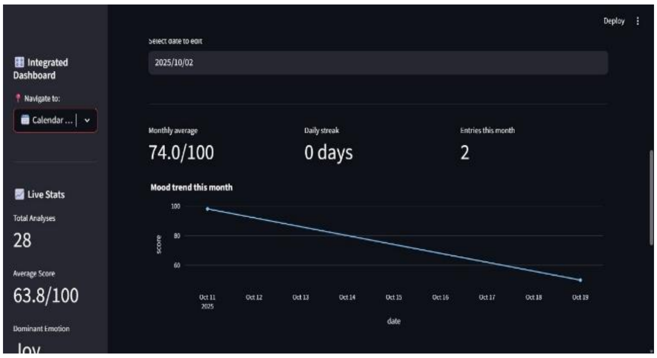

Fig 14 – Outcome in Graphical form

5. Conclusion and Future Work

By utilizing both Deepfaces (facial recognition software) and BERT (text analysis), we have successfully created an integrated interface for providing emotional wellbeing insight using artificial intelligence-based sentiment/emotion identification technology. This combined approach provides greater emotional insight than simply identifying whether something is "positive" or "negative." Users will be able to gain a greater understanding of their emotional state as well as provide opportunity for improved responses to their emotional states.

Additionally, the platform’s ability to store emotional history, recognize emotional patterns over time, and deliver personalized wellbeing recommendations adds to its effectiveness. Future additions will allow for voice emotion detection, use of data from wearables, and the introduction of more advanced algorithms. With these additions, this technology will provide an increasingly accurate, efficient, and applicable solution toward developing an all-inclusive artificial intelligence-based approach for providing assistance in maintaining one’s own mental well-being.

Future scope

- **Chatbot** for general purpose use and to resolve some questions of the user related to their emotions and related activities.
- **Emergency alert system** for guardians or counselors, ensuring proactive care in critical emotional states.
- **Mobile application deployment** for greater accessibility and real-time use by a broader audience.

References

1. M. Soleymani, M. Pantic, and T. Pun, “Multimodal emotion recognition in response to videos,” *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 211–223, 2012.

2. R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 18–37, 2010.
3. Y. Wu et al., "M GAT: Multi-Granularity Attention Based Transformers for Multi-Modal Emotion Recognition," *IEEE Trans. Affective Comput.*, 2024.
4. M. Khan et al., "MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion," *Scientific Reports*, vol. 15, 2025.
5. A. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. NAACL-HLT*, 2019.
6. M. H. Yi et al., "HyFusER: Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention," *Applied Sciences*, vol. 15, 2025.
7. S. Hazmoune et al., "Using transformers for multimodal emotion recognition," *Eng. Appl. Artif. Intell.*, vol. 108, 2024.
8. F. Zeng et al., "Multimodal temporal fusion for emotion recognition," *IEEE Access*, 2023.
9. S. Poria et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *Proc. ACL*, 2019.
10. H. Bhoite, "Real-Time Multimodal Emotion Recognition for Edge Virtual Assistants Using Lightweight Transformer Models," *TechRxiv*, 2025.
11. J. Liu et al., "Cross-modal transformers for wild emotion recognition," *IEEE Trans. Multimedia*, 2024.
12. K. Singh et al., "TACFN: Adaptive cross-modal networks for emotion recognition," *IEEE Trans. Neural Networks Learn. Syst.*, 2023.
13. L. Wu et al., "Benchmarking Deep Facial Expression Recognition," *arXiv preprint arXiv:2311.02910*, 2023.
14. P. Zhang et al., "Physiological signal integration in affective computing," *IEEE Trans. Affective Comput.*, 2024.
15. Usha Kosarkar, Gopal Sakarkar (2024), "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis", Journal of Multimedia Tools and Applications, 1380-7501, https://doi.org/10.1007/s11042-024-19220-w
16. Anupam Chaube, Usha Kosarkar "Enhanced Deep Learning-Based Feature Analysis for Copy-Move Forgery Detection", 2024 Journal of Information Systems Engineering and Management,9(4s) e-ISSN:2468-4376, https://www.jisem-journal.com/