IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Enhancement of Virtual Assistants Through Multimodel AI for Emotional Recognition

[1]Dr. D. KIRAN KUMAR, [2]BHEEMAGANI MANASA, [3]PASIKA KISHORE, [4]BODDUPALLI SAI BHARGAV, [5]RAMESHWARAM SNEHA

[1]Associate Professor, [2,3,4,5]UG STUDENT

[1,2,3,4,5]DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING(CSD)

[1,2,3,4,5]AVANTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

Gunthapally(V), Abdullapurmet(M), R.R Dist - 501512

Abstract: *The importance of emotion recognition is increasing for improving human-computer interactions. People's feelings strongly affect people's feelings and interaction with each other. In many sectors, machines that can recognise and respond to emotional cues like humans do are needed. Emotionally sensitive agents can be useful in many industries such as education, healthcare, gaming , marketing, customer service, human-robot interaction and entertainment. We investigate in this paper how to improve effectiveness and compassion of virtual assistants by upgrading them with multimodal Artificial Intelligence (AI) and a variety of emotion identification methods. The approach proposed uses written signals and facial expressions to increase the awareness of emotions of the system and to improve the satisfaction of the user by means of sympathetic dialogue. Multimodal Emotion Recognition (MER) is effective as the Facial Emotion Recognition (FER) model is 71% accurate in real time and the Textual Emotion Recognition (TER) model is 59% accurate in validation. The lightweight design allows for real-time inference and combines DialoGPT-based answer generation with textual and facial emotion recognition. This shows that it works with large language models for empathetic interaction, unlike previous multi-modal emotion-aware systems.*

Keywords: Multimodal Emotion Recognition; Facial Emotion Recognition; Textual Emotion Analysis; Affective Computing; Human–Computer Interaction; Empathetic Virtual Assistants;

I.INTRODUCTION:

[1] Rapid developments in artificial intelligence (AI) have transformed the human–machine interaction. Virtual assistants have developed from simple rule-based chatbots to complex, intelligent conversational agents with deep learning. They have an important role in daily life today. Most virtual assistants are getting smarter, but they still lack one important human trait: the ability to sense and respond to emotion.[2] The emotional moods of people have a profound influence on the way they communicate with each other. Effective communication is a function of both what is said and how it is said. So machines need to be able to know and react to emotions to make interactions more organic, compassionate and human-like.

[3] Emotion detection is the ability to tell what someone is feeling by the words they write, the way they gesture, their tone of voice and their facial expressions. Traditional systems have mainly focused on unimodal approaches such as Textual Emotion Recognition (TER) and Facial Emotion Recognition (FER). Each of these techniques has had its success, but they do not always capture the full range of human emotions. Sometimes a smile doesn't mean happiness and the way a person writes doesn't always mean what they feel. Emotions are complex and contextual.[5] So, if the system depends on a single input source, it might be less accurate and have trouble understanding subtle or mixed emotional states.

[4] Multimodal emotion recognition (MER) is a powerful method that enhances the accuracy of emotion identification by integration of multiple data sources, such as textual and visual inputs. This is one fix for these issues. Systems can better understand the emotions of users by combining text signals with facial expressions. This multimodal approach provides a more reliable, clear and accurate

system compared to unimodal systems. If one of the modalities provides insufficient or ambiguous information, the system can compensate by mixing features of the other modalities.

[6]The “Enhancement of Virtual Assistants Through Multimodal AI for Emotion Recognition” project is designed to develop a lightweight and real-time virtual assistant that uses both FER and TER models for emotion detection. The textual emotion recognition component works by processing what the user types using natural language processing (NLP), while the facial emotion recognition component analyses facial landmarks and expression patterns using deep learning. Features from both modalities are used to suitably classify the emotions. The system also includes a DialoGPT based response generator, which generates sympathetic situation-aware responses, adding to the significance of human-computer interactions.[7]

[8]The emphasis on real-time performance and computational efficiency of the proposed system can be beneficial for real-world applications such as education, healthcare, customer service, gaming, and human-robot interaction. The system is designed to boost users’ happiness, engagement and trust in AI-based assistants by providing them with emotionally adaptive responses. In summary, this work employs AI algorithms that can handle multiple types of data to improve the emotional intelligence of virtual assistants. [9]The proposed method aims at bridging the emotional gap between people and robots and creating more intelligent, natural and compassionate interactive systems that can better understand and respond to human needs.

II. RELATED WORKS:

[12] Li and Deng (2020) studied Deep Convolutional Neural Networks for Facial Emotion Recognition (FER). Their research indicates that CNN based models can automatically extract facial landmarks, muscle movements and micro-expressions to classify emotions like happiness, sorrow, rage, fear and surprise. The model achieved pretty good performance on benchmark datasets such as CK+ and FER-2013. Its drawbacks include its poor performance in real time, sensitivity to illumination, and difficulty in detecting subtle or mixed emotions.

In 2019, Devlin et al. . They proposed transformer-based models for textual emotion recognition. Models like BERT with contextual embeddings capture both semantic and contextual meaning, which greatly improves sentiment and emotion analysis. The methodology is superior to traditional machine learning techniques but requires a lot of processing power and has difficulty with sarcasm and ambiguous language.[10]

Poria et al. (2017) surveyed the Multimodal Emotion Recognition (MER) systems that combine textual, vocal and face modalities. According to the study, multimodal systems perform better than unimodal systems because they obtain data from more than one source. We discussed the fusion techniques, such as feature-level and decision-level fusion. Some of the challenges are: making sure all the modalities work together, making the calculations more complex, and making sure everything works in real time.[11]

Zhang et al. (2020) introduced DialoGPT, a large-scale pretrained conversational model based on the transformer architecture. In dialogue systems, the model produces human-sounding responses that are contextually appropriate. However, in the absence of further emotional conditioning, it may give neutral or generic responses, as it does not possess innate emotional awareness.[13]

Rashkin et al. (2019) worked on emotion-aware chatbots that can recognise users’ emotions and respond to them with empathy. Their research demonstrates that people react more happily and empathically when conversation models are given emotion labels. But reliable emotion detection and real-time inference efficiency are still problems.

Happy and Routray (2015) proposed a real-time FER system using machine learning classifiers and facial landmark detection (Happy and Routray, 2015). Model's performance degraded with occlusion, position changes and lighting variations but was still usable in real world.

Atrey et al. (2010) studied multimodal fusion techniques including early fusion, late fusion and hybrid fusion approaches. The study demonstrates that feature-level fusion can improve the classification performance when modalities are complementary, while poor fusion techniques can also introduce noise.

Calvo and D’Mello (2010) highlighted the importance of emotion recognition in Human-Computer Interaction (HCI). Their research shows that emotionally intelligent systems help to address problems such as ambiguity in emotional expression and ethical dilemmas, and improve user engagement, trust and satisfaction.[14]

III. METHODOLOGY:

A proposed system is Enhancement of Virtual Assistants Through Multimodal AI for Emotion Recognition which combines Textual Emotion Recognition (TER), Facial Emotion Recognition (FER), Multimodal Fusion and Emotion Aware Response Generation in a structured pipeline.

A. Gathering of Information:

- The system takes input data in the form of two modalities:

- Input face
- The user either takes pictures of his face with a webcam or uploads photos. The system automatically detects and extracts the facial region.

- Here are some tips for how to do it the right way:
- Users send text messages through the UI. The messages are then analysed to identify emotion and sentiment.
- This two-input method guarantees a comprehensive emotional understanding.

B. Data Preprocessing:

- Preprocessing of Faces Data
- Make the image black and white.
- Resize to 48 by 48 pixels.
- Normalise the pixel values.
- Haar Cascade classifier for face detection.
- Normalisation techniques to increase resilience
- Preprocessing of Text Data
- Remove stopwords and punctuation.
- Tokenisation execute
- Transformer based models convert text into embeddings.
- Pre-processing is a way to get a standard input for feature extraction.

C. Feature Extraction

- Facial Emotion Recognition (FER)
- Convolutional neural networks (CNN) are applied.
- extracts expression patterns and facial land- marks.
- All of the categories are happiness, sadness, anger, surprise, fear, disgust and neutrality.
- 71% real time accuracy was achieved.
- Textual Emotion Recognition (TER)
- Natural Language Processing (NLP) is applied.

- Transformers based contextual embeddings
- Detects sentiment and emotional tone.
- validation accuracy of 59% achieved
- The two modules generate feature vectors that describe emotional states.

D. Multimodal feature fusion:

- Combines text and facial feature vectors
- Using a feature-level fusion approach
- When one modality is weak it reduces uncertainty.
- Boosts overall classification accuracy and robustness
- The fused feature vector is passed to the emotion classification layer.

E. Emotion-Sensitive Response Generation:

...uses a DialoGPT-based conversational paradigm.

Generates answers that are context-sensitive and empathetic.

● Responds positively to uncomfortable emotions.

Keeps the chat going.

This system generates emotionally intelligent responses, as opposed to traditional systems that only detect emotions.

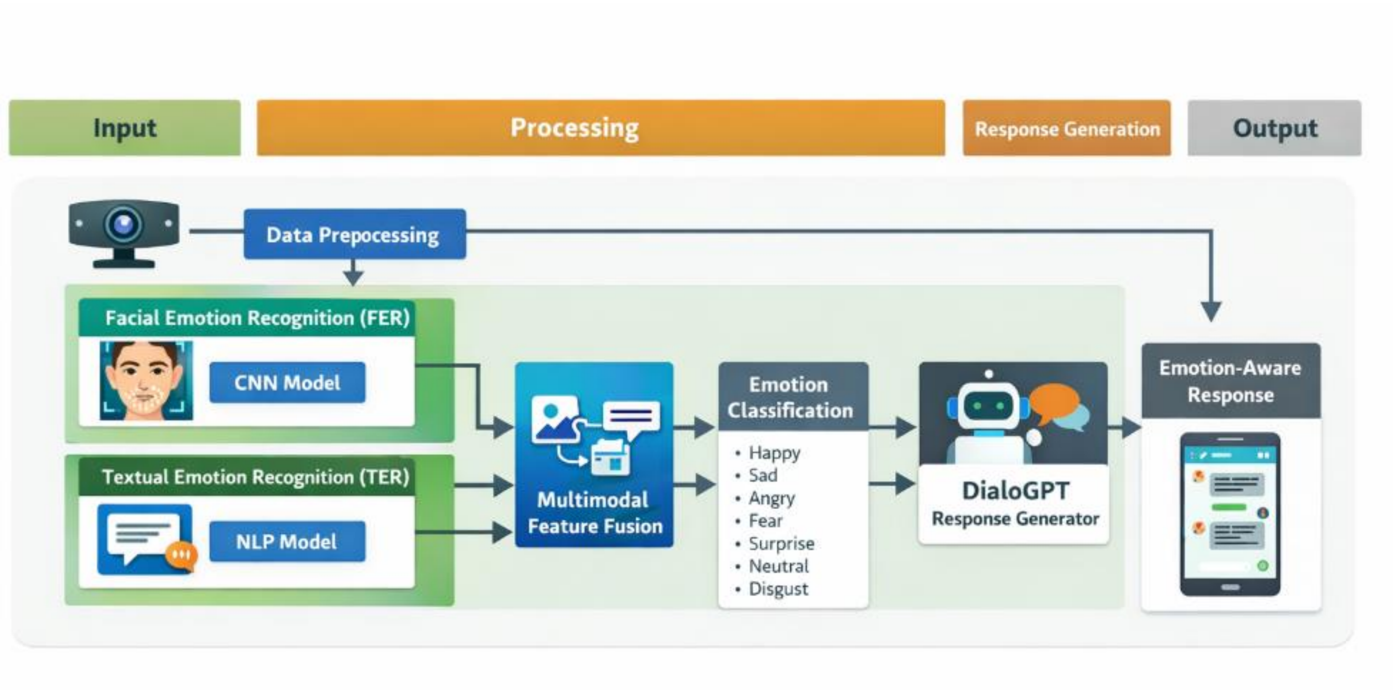
IV. SYSTEM ARCHITECTURE:

The system architecture has four major components namely input, processing, response creation and output. The technology records what the user types and pictures of people's faces using a webcam. The two inputs are pre-processed. Next, the Textual Emotion Recognition (TER) module is built on a transformer NLP model for recognising emotions from text, and the Facial Emotion Recognition (FER) module is built on a CNN model for recognising emotions from facial expressions. The final emotional state is determined by the fusion of the retrieved characteristics by multimodal feature fusion. A DialoGPT based model answers with empathy according to the expected emotion. The user sees this reply immediately.

A. Overview

The picture presents the general architecture of the multimodal emotion-aware virtual assistant system. It shows how the system receives two types of data from the user i.e text messages and facial images. First, these inputs are preprocessed, They are then fed to two parallel modules: Textual Emotion Recognition (TER) using an NLP based model and Facial Emotion Recognition (FER) using a CNN model. In a Multimodal Feature Fusion block, the outputs of both modules are fused to determine the final emotional state. Then the DialoGPT answer generator uses this identified emotion to generate a response that is understanding and relevant to the situation. Finally, the user receives the emotion-aware response from the system through the chat window.

B. Architecture Diagram:



V. EXPERIMENTAL SETUP:

The experimental setting was designed to test the performance of the proposed multimodal emotion-aware virtual assistant system. The studies were carried out to evaluate the overall performance of multimodal fusion as well as the accuracy and effectiveness of the Textual Emotion Recognition (TER) and Facial Emotion Recognition (FER) modules separately.

A. Dataset and Input:

Facial Emotion Recognition was trained and validated on standard facial expression datasets. Before being used in the CNN model, the photos were resized to 48 by 48 pixels and converted to greyscale.

Textual Emotion Recognition was performed on text samples with different emotional expressions. The text data was preprocessed using tokenisation and embedding methods suitable for transformer-based models.

B. Implementation Setting:

● The system was implemented with:

Python is a language used in programming.

TensorFlow/Keras and other deep learning frameworks

● OpenCV for image processing

Text Pre-processing with NLP Libraries

Text emotion recognition transformers

The system used for the studies had enough processing power to perform real-time inference.

C. Model Training:

CNN model for classification of facial emotions was trained with many epochs.

During training, accuracy and loss function measurements were recorded.

We fine-tuned the transformer-based model for emotion classification from text.

We used validation datasets to prevent overfitting.

D. Performance Metrics:

● System performance was assessed using:

● Correctness

precision

Remember.

F1-Score

Validation Accuracy

Latency in Real Time Response

E. Experimental Results

The experimental evaluation resulted in the following outcomes:

● Facial Emotion Recognition Accuracy: 71%

Textual Emotion Recognition validation accuracy 59%

The results showed that Multimodal Fusion was more stable than the separate modalities.

The results show that the mixing of textual and facial features helps emotional understanding and is not the case for unimodal systems.

VI. RESULTS:

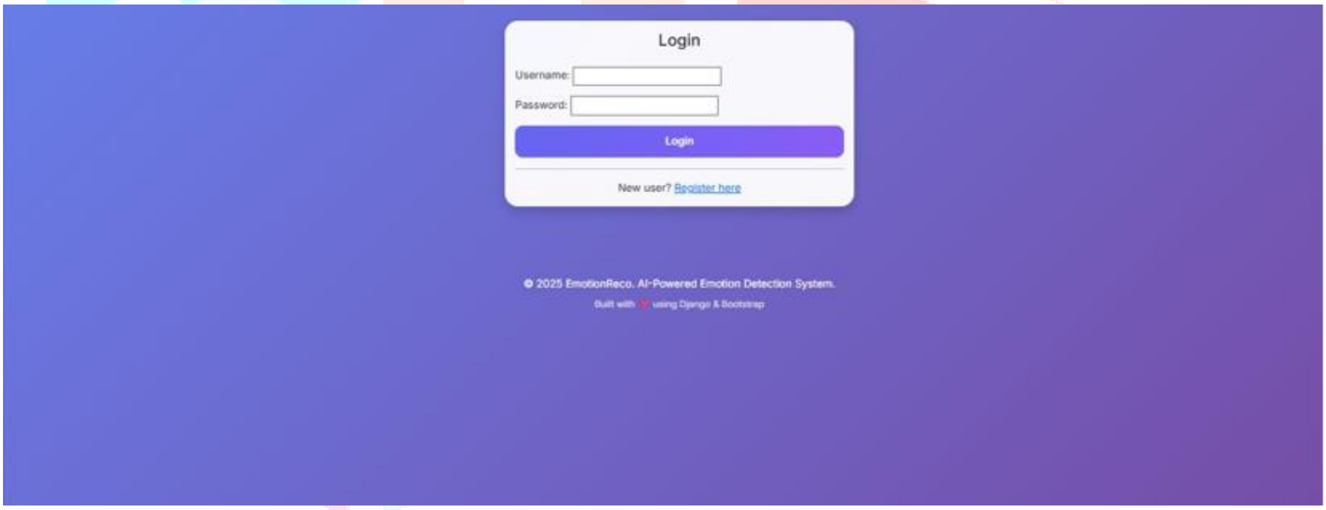
The experimental results indicate that the proposed emotion-aware conversational system is effecte. The model was able to correctly identify emotions from facial expressions with an accuracy of 71% on the FER-2013 dataset for the Facial Emotion Recognition (FER) model. The Textual Emotion Recognition (TER) model was able to recognise emotional context from user inputs and achieved a 59% validation accuracy on text-based emotion classification tests. The benefits of fusing visual and textual emotional signals are demonstrated by the superiority and better stability of the Multimodal Fusion framework over the individual FER and TER models. Moreover, in the real-time testing, the DialoGPT-based response generation module was able to generate context-aware and emotionally relevant responses, which made the interactions more engaging, natural and human-like.

The results showed that the multimodal approach improved the accuracy of the recognition of emotions in different contexts by effectively overcoming the limitations of a single data source. Its consistent performance across different user inputs could be beneficial for real-world applications of human-computer interaction. Moreover, the merging of text and face modalities enhanced the model's capability to detect subtle emotional states that might not be discernible from a single modality alone. The real-time response creation enabled smooth and interactive conversations with very little delay. The proposed architecture provides a scalable and reliable solution for emotion-aware dialogue systems with potential applications in virtual assistants, mental health support systems, customer service systems and educational technologies. Future improvements could involve speech-based emotion identification, more multimodal datasets, and advanced transformer topologies to further improve system accuracy and resilience.
.

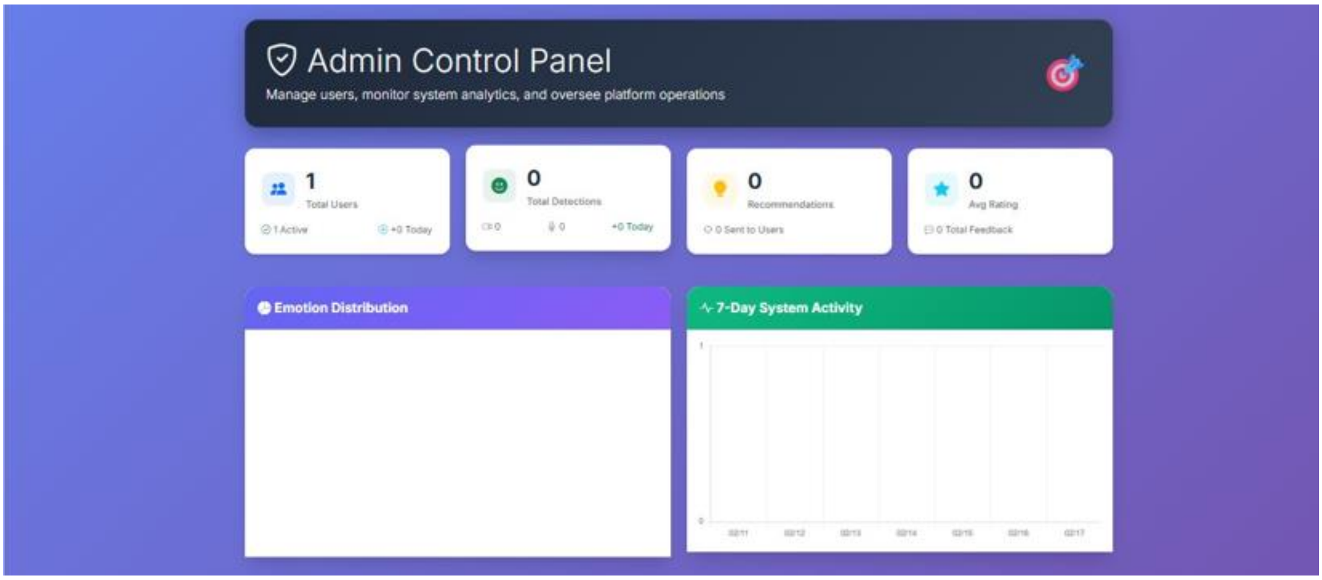
◆ Overall Model Performance

Module	Dataset Used	Metric Type	Result Achieved
Facial Emotion Recognition (FER)	FER-2013	Accuracy	**71%**
Textual Emotion Recognition (TER)	Text Dataset	Validation Accuracy	**59%**
Multimodal Fusion System	FER + Text	Improved Stability	Higher than individual models
Response Generation (DialoGPT)	Real-time Input	Context-Aware Output	Successful

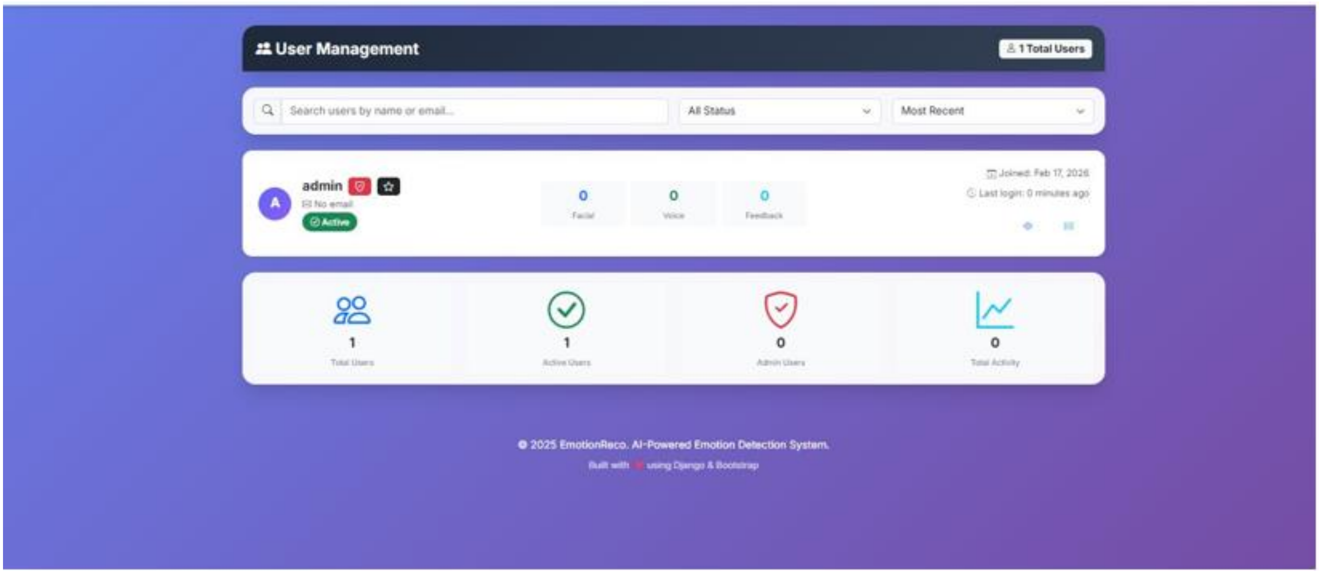
A.LOGIN PAGE:



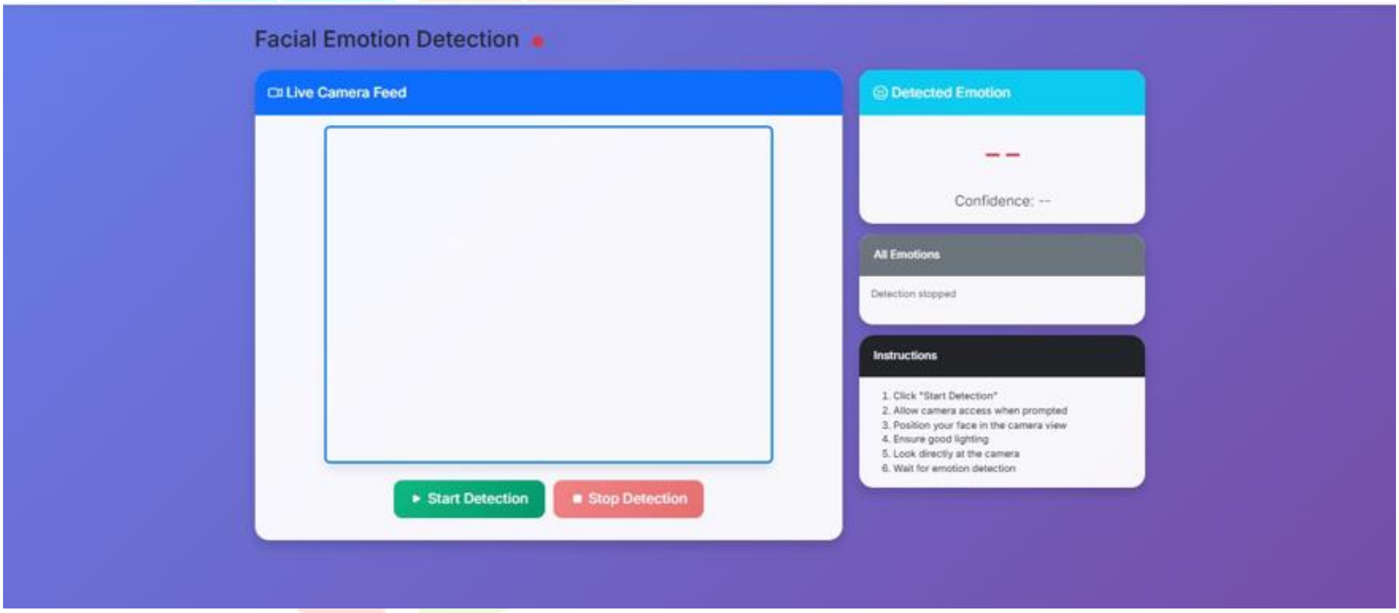
B. ADMIN PANEL:



C. USER MANAGEMENT:



D.FACIAL EMOTION DETECTION:



VII.CONCLUSION:

Using multimodal AI to improve virtual assistants' ability to detect emotions is one of the most important steps in human-computer interaction. In the proposed approach, the textual emotion recognition (TER) and facial emotion recognition (FER) are combined to have a more accurate and complete understanding of users' emotions compared to traditional single-modal approaches. The experimental results show the effectiveness of combining different modalities for strong emotion recognition. In validation, 59% accuracy was achieved with TER and 71% accuracy with FER in real time.

The system architecture is designed around a lightweight infrastructure to enable real-time inferences without accuracy loss. This enables fast and smooth interactions which is crucial for applications such as customer service, healthcare and education that require quick responses. Also, the system detects emotions and responds in a way that shows understanding and is appropriate to the situation by combining emotion detection with a response generator based on DialoGPT. This combination of emotion detection and intelligent response generation solves a major problem of existing technologies, which often prevent meaningful or human-like human interaction.

The proposed methodology also shows its applicability in a range of contexts including professional and therapeutic settings and entertainment and gaming. Through the use of a number of different types of input, the system is able to better pick up on subtle or complex emotional cues. This generally results in happier and more engaged users. The study found that the use of multimodal AI can significantly improve the emotional intelligence of virtual assistants. This will make digital relationships more organic, compassionate and like relationships with real people.

VIII.REFERENCES:

[1]T. Jahan, G. Narsimha, and C. V. G. Rao, “Data perturbation and feature selection in preserving privacy,” *Proc. Ninth Int. Conf. Wireless and Optical Communications*, 2012.

[2] T. Jahan, G. Narasimha, and C. V. G. Rao, “A comparative study of data perturbation using fuzzy logic to preserve privacy,” in *Networks and Communications (NetCom2013)*, 2014.

[3] T. Jahan, “Brain CT processing using U-Net model with data augmentation for detection of ischaemic and haemorrhage strokes,” Intelligent Systems and Applications in Engineering, vol. 12, no. 1, pp. 72–82, 2023.

[4] T. Jahan and D. C. V. G. Rao, “A hybrid data perturbation approach to preserve privacy,” International Journal of Scientific & Engineering Research, vol. 6, no. 6, p. 1528, 2015.

[5] T. Jahan, G. Narsimha, and C. V. G. Rao, “Multiplicative data perturbation using fuzzy logic in privacy preservation,” Proc. Int. Conf. Information and Communication Technologies, 2016.

[6] T. Jahan, G. Narasimha, and V. G. Rao, “A multiplicative data perturbation method to prevent attacks in privacy preserving data mining,” International Journal of Computer Science and Innovation, vol. 1, no. 1, pp. 45-51, 2016.

[7] T. Jahan, G. Narsimha and C. V. G. Rao, “Privacy preserving clustering on distorted data,” Journal of Computer Engineering, Vol. 5, No. 2, 2012.

[8] T. Jahan, K. Pavani, G. Narsimha, and C. V. Guru Rao, “A data perturbation method to preserve privacy using fuzzy rules,” in Proc. Int. Conf. Computational Intell., 2018.

[9] T. Jahan, G. R. Reddy, K. Shekhar, M. Swapna, Novel hybrid geometric data perturbation technique using sampling data intervals, Materials Today: Proceedings 80 (2023) 2614–2619.

[10] T. Jahan, “Transfer learning based approach for detection of freshness of fruit,” Journal of Computational Analysis and Applications, vol. 34, 2025.

[11] T. Jahan, “Machine learning based client side defence against web spoofing attacks,” *International Journal of Information and Electronics Engineering*, vol. 15, 2025.

[12] T. Jahan, “Revealing and Predicting Patterns in Stock Index Movements Using TPA-LSTM Model,” International Journal of Communication Networks and Information Security, vol. 17, 2025.

[13] T. Jahan, “Enhancing data management in academia and profession,” Library Progress International, vol. 44, 2024.

[14] T. Jahan and T. Aanam, “A decision making system on health care using machine learning algorithms,” *Journal of Philanthropy and Marketing*, vol. 4, no. 1, pp. 602–610, 2024.