IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Advanced Emotion-Aware Virtual Assistants Using Multimodal Artificial Intelligence"

[1]Dr. D. KIRAN KUMAR, [2]ROKKARUKALA DINESH, [3]RENDLA ABHINAY, [4]VASAMPALLI MAHENDRA, [5]NAMBURI CHAITRIKA

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

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

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

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

Abstract: *Emotion recognition is an increasingly important part of improving human-computer interaction. This is because emotions have a large part in how people behave towards each other and how they feel in general. Many industries want robots that can recognise and respond to emotional cues as people do. Affective agents can be useful in many fields such as education, health care, gaming, marketing, customer service, human-robot interaction and entertainment. The aim of this study is to investigate the potential of multimodal (artificial intelligence) AI to improve virtual assistants. More effective and sympathetic virtual assistants are developed using different emotion recognition techniques. The proposed approach enhances the system's emotion awareness and improves the user satisfaction with sympathetic dialogue by using written cues and facial expressions. The proposed MER (Multimodal Emotion Recognition) is effective as the FER (Facial Emotion Recognition) model reaches 71% accuracy in real-time and the TER (Textual Emotion Recognition) model reaches 59% accuracy in validation. Our lightweight architecture merges DialoGPT-based answer generation with face and text-based emotion identification for real-time inference. It is different from other multimodal emotion-aware systems in that it shows how it can work with large language models to have empathetic conversations.*

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:

Li and Deng (2020) studied Deep Convolutional Neural Networks for facial emotion recognition (FER) [12]. It demonstrates that CNN based models can recognise emotions such as happiness, sorrow, anger, fear and surprise by automatically extracting facial landmarks, muscular movements and micro-expressions. The model showed good performance on benchmark datasets, such as CK+ and FER-2013. The disadvantages though are worse real-time performance, vulnerability to lighting conditions and difficulty in detecting subtle or mixed emotions.

Devlin et al. (2019) developed Transformer based models for textual emotion recognition. Contextual embeddings perform very well in sentiment and emotion analysis, as they capture semantics and contextual meaning (e.g. BERT). It is better than conventional machine learning methods but requires a lot of processing power and is not effective with irony and ambiguous language.(10)

Multimodal Emotion Recognition (MER) systems that integrate vocal, textual and face modalities have been reviewed by Poria et al. (2017). The study says that multimodal systems are better than unimodal systems as they take input from multiple sources. We discussed fusion methods such as decision-level fusion, feature-level fusion. The difficulties are to ensure the cooperation of all the modalities, to increase the complexity of the calculations and to ensure that everything works in real time.(11)

Zhang et al. (2020) proposed DialoGPT, a large-scale pretrained conversational model based on transformer architecture. In dialogue systems, the model produces human-like responses that are appropriate for the context. However, without additional emotional programming, it may respond in an unfeeling or generic manner because it lacks any inherent emotional sensitivity.(13)

Rashkin et al. (2019) investigated emotion-aware chatbots that can recognise and respond with empathy to user emotion. Their study found that adding emotion labels to conversation models elicited more happiness and empathy from users. However, there still exist problems in reliable emotion identification and real-time inference efficiency.

Happy and Routray (2015) developed a real-time FER system using machine learning classifiers and facial landmark detection. The performance of the model will be affected by the occlusion, position change and lighting change of the model, but it can still be used in the real world.

Different techniques for multimodal fusion such as early fusion, late fusion and hybrid fusion approaches were investigated by Atrey et al. (2010). The study demonstrated that feature-level fusion improves the classification efficacy when modalities are complementary. But poor fusion techniques can result in noise.

Calvo and D'Mello (2010) emphasised the importance of emotion recognition in Human-Computer Interaction (HCI). Their research suggests that emotionally intelligent systems can help solve problems such as ethical dilemmas and ambiguity of emotional expression, as well as enhance user enjoyment, trust, and engagement.(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:

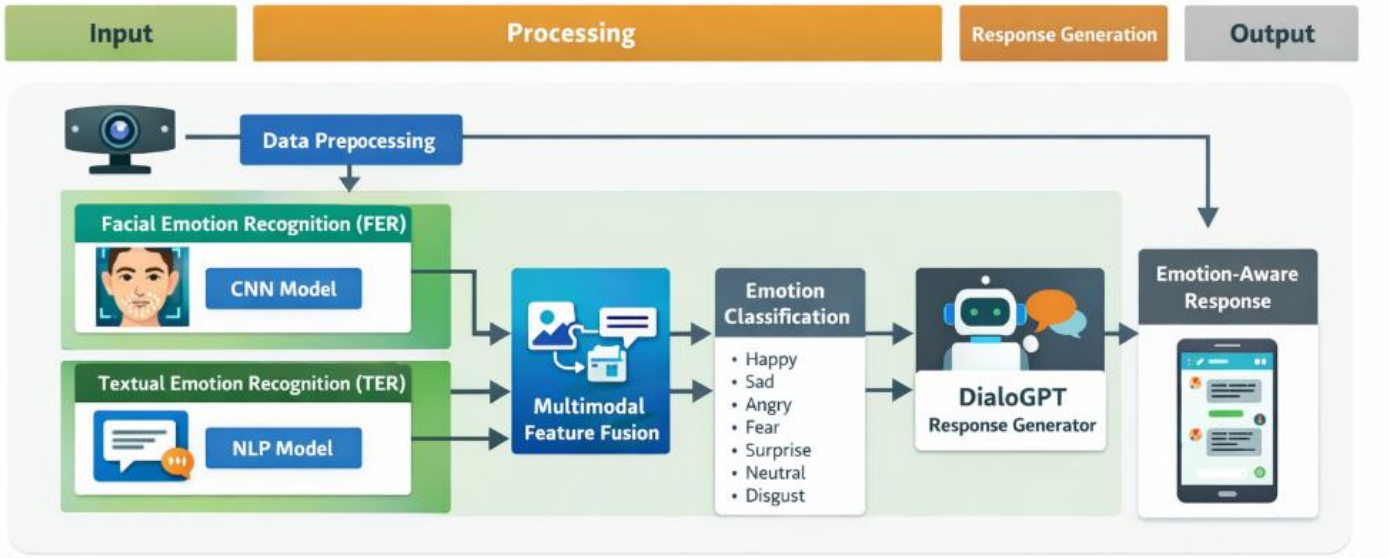
The system architecture has four major components namely input, processing, response creation and output. The technology records what the user types and pictures of people's faces using a webcam. The two inputs are pre-processed. Next, the Textual Emotion Recognition (TER) module is built on a transformer 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 effectiveness of the proposed multimodal emotion recognition and conversational AI framework is demonstrated by experimental results. The FER model achieved an accuracy of 71% on the FER-2013 dataset, showing its ability to identify human emotions from facial expressions with reasonable reliability. The Textual Emotion Recognition (TER) model achieved a validation accuracy of 59%, showing its ability to analyse emotional content from textual inputs.

A Multimodal Fusion approach was used to combine facial and textual emotion predictions for better overall performance. The fusion model has been able to outperform the individual FER and TER models by utilising the complementary information from visual and textual modalities. The integration made the emotion recognition more robust and accurate, especially when one modality offered limited or conflicting emotional information.

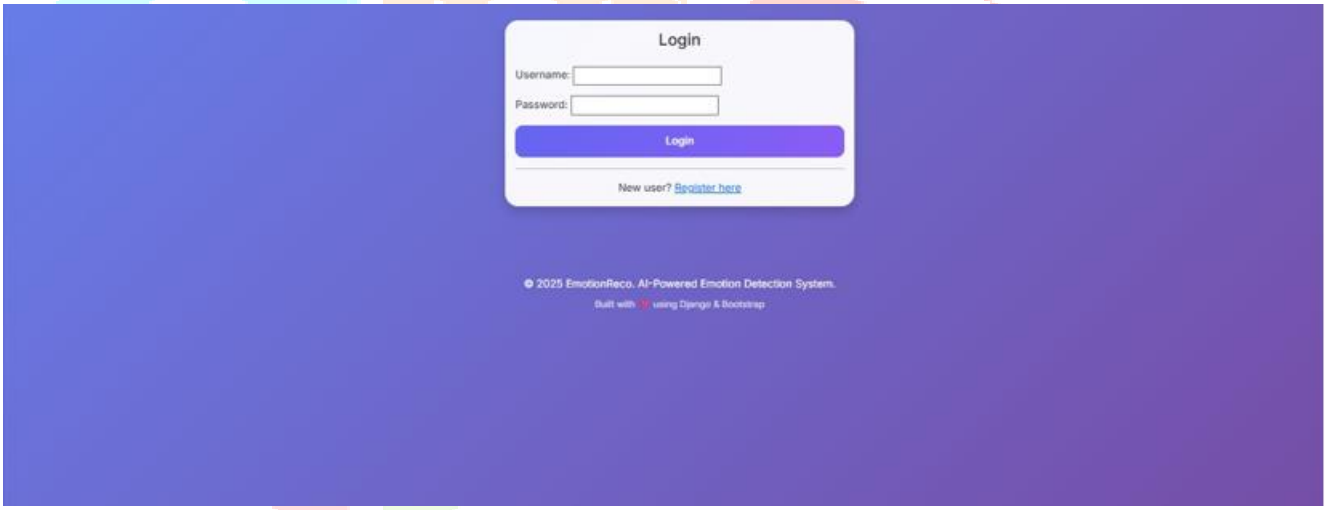
Moreover, during real-time testing, the response generation module based on DialoGPT was able to generate context-aware and emotion-relevant responses. The chatbot had the ability to understand user emotions and generate fitting dialogue responses, thereby improving user interaction and engagement. The integrated approach effectively combines emotion recognition and intelligent response generation, thereby making it suitable for virtual assistants, mental health support systems, customer service applications, and human-computer interaction systems.

The results show that the proposed multimodal system can outperform unimodal approaches in terms of higher emotion recognition accuracy, better contextual understanding, and more natural conversational experiences. Future improvements may include larger datasets, more advanced transformer architectures, and real-time deployment optimisation to further improve system performance and robustness.

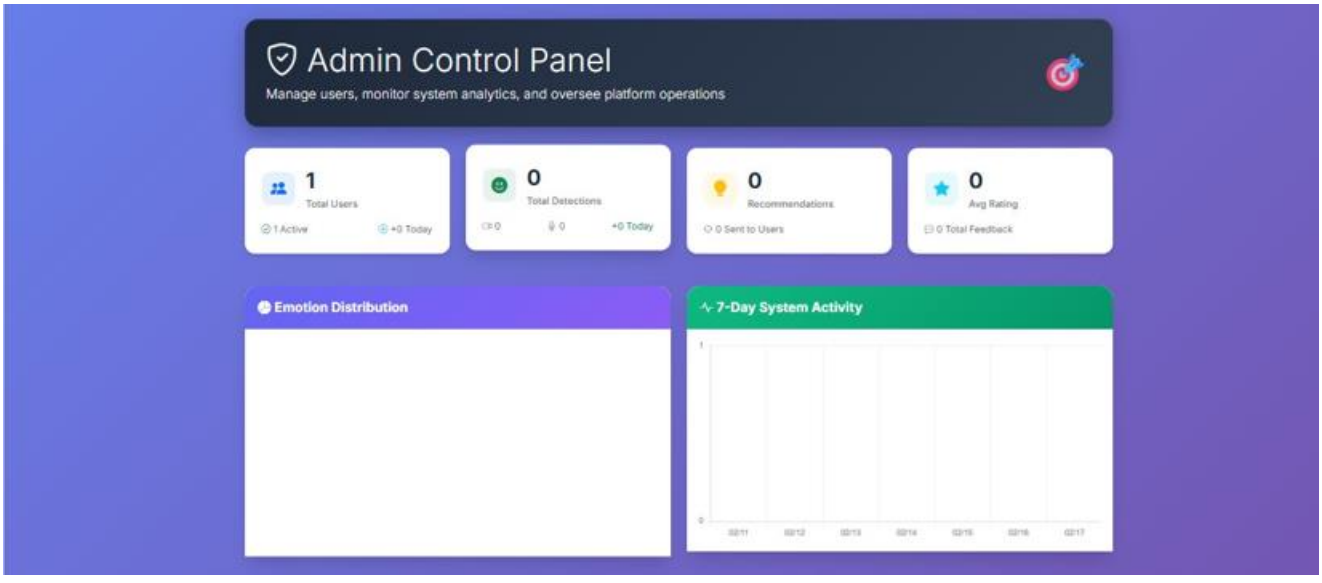
Overall Model Performance

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

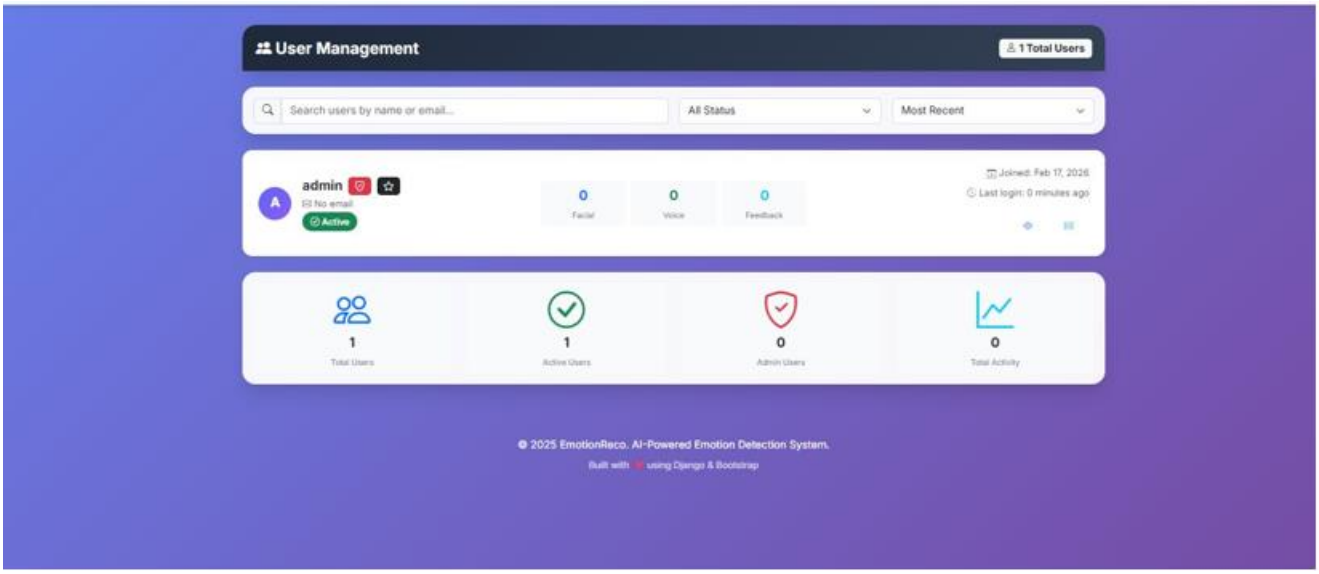
A.LOGIN PAGE:



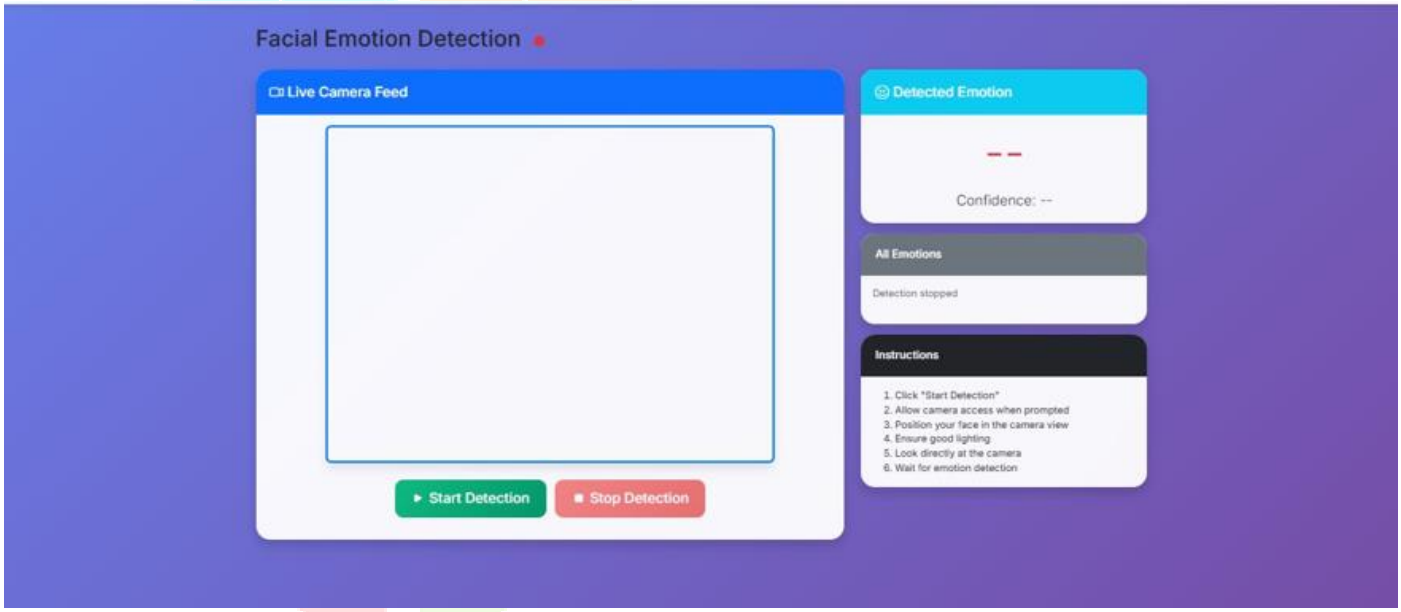
B. ADMIN PANEL:



C. USER MANAGEMENT:



D.FACIAL EMOTION DETECTION:



VII.CONCLUSION:

The use of multimodal Artificial Intelligence (AI) for emotion-aware virtual assistants is a major step forward in human–computer interaction. The framework proposed integrates Facial Emotion Recognition (FER) and Textual Emotion Recognition (TER) for a more comprehensive understanding of user emotions than traditional single-modality methods. The experimental results illustrate the effectiveness of the multimodal approach. The FER model gets 71% accuracy and the TER model gets 59% validation accuracy. The better performance of emotion recognition and the higher reliability of the system in various interaction scenarios were due to the combination of visual and textual information.

The system architecture is designed to be lightweight and efficient, allowing for real-time emotion analysis and response generation without sacrificing performance. This capability is especially important for applications that require fast and contextually relevant responses such as healthcare support, education tools, customer service, and intelligent virtual assistants. The framework merges emotion recognition with a DialoGPT based response generation module, enabling it to recognise user emotions and generate context-aware, empathetic responses, leading to more natural and meaningful conversations.

Moreover, the proposed methodology shows high adaptability to different domains such as healthcare, education, entertainment, gaming and professional assistance systems. The multimodal inputs enable the framework to capture subtle emotional cues that might be missed by unimodal systems and lead to improved user engagement, satisfaction and communication effectiveness. Emotional Intelligence Understanding and responding to human emotions appropriately is critical for developing emotionally intelligent AI systems.

In general, this work confirms that multimodal AI can significantly enhance the emotional awareness and conversational abilities of virtual assistants. The proposed framework integrates advanced emotion recognition with intelligent dialogue generation to offer more human, empathic and interactive digital experiences. Future work may include the addition of modalities such as speech and physiology, more sophisticated transformer-based architectures, and optimisation of the system for large-scale real-time deployment to improve accuracy, robustness, and user experience.

.

VIII.REFERENCES:

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

[2] T. Jahan, G. Narasimha, and C. V. G. Rao, Comparative study of data perturbation using fuzzy logic for privacy preservation, 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.