IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2025, accepted 3 June 2025, date of publication 9 June 2025, date of current version 18 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3577664

RESEARCH ARTICLE

Enhancement of Virtual Assistants Through Multimodal AI for Emotion Recognition

SHAUN GEORGE RAJESH, SMRITI VIPIN MADANGARLI, GAURI SANTOSH PISHARADY, AND ROLLA SUBRAHMANYAM

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India

Corresponding author: Rolla Subrahmanyam (rolla.subrahmanyam@vit.ac.in)

ABSTRACT Emotion recognition is becoming increasingly critical for enhancing human-computer interactions, as emotions play a vital role in shaping human interactions and overall well-being. Machines that can detect and respond to emotional cues similar to humans are essential in multiple industries. Emotionally responsive agents find applications in education, healthcare, gaming, marketing, customer service, human-robot interaction, and entertainment. This study explores the potential of enhancing virtual assistants through multimodal Artificial Intelligence (AI), utilizing various emotion recognition techniques to create more empathetic and effective systems. The proposed methodology makes use of facial expressions and textual cues to enhance the emotional awareness of the system and achieve user satisfaction through empathetic conversation. The Facial Emotion Recognition (FER) model achieved 71% real-time accuracy, whereas the Textual Emotion Recognition (TER) model achieved 59% validation accuracy, demonstrating effective Multimodal Emotion Recognition (MER). Unlike prior multimodal emotion-aware systems, our lightweight architecture ensures real-time inference and uniquely integrates facial and textual emotion recognition with DialoGPT-based response generation — demonstrating compatibility with large language models for empathetic dialogue.

INDEX TERMS Virtual assistants, large language models, facial emotion recognition, BERT, computer vision, neural networks.

I. INTRODUCTION

Emotions play a crucial role in human communication by influencing how individuals interact and respond to various social contexts. As conversational agents become more integrated into daily life, enhancing their ability to understand and respond to human emotions is essential to create more engaging, empathetic, and effective interactions. Traditional text-based models for emotion recognition often overlook the rich emotional cues present in non verbal communication and often struggle with fuzzy boundaries between emotions [1]. Complex structures, such as idioms and sarcasm, are often misinterpreted resulting in incorrect emotions being identified. On the other hand, systems that focus solely on FER may miss the nuances conveyed by language. These systems also falter in real-time environments because of the lack of diversity in emotion portrayal and facial positioning within datasets [2]. This indicates that single-modality systems often fail to capture complex emotional cues. A review of the literature revealed that many researchers have highlighted how incorporating multiple modalities can complement one another, providing a more comprehensive understanding of emotion. Hence, integrating FER and TER could enable richer emotional understanding by combining visual and textual data.

This paper proposes an innovative methodology that integrates both FER and TER for creating an emotion-aware virtual assistant (Fig 1). The system leverages the FER model based on the Vision Transformer (ViT) architecture specifically, ViT-Tiny, trained on the AffectNet dataset, to capture real-time facial expressions. A MiniLM-based TER model is utilized to analyze the user's textual input, categorizing it into one of seven classes of emotions: joy, sadness, anger, surprise, fear, disgust, and neutral.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

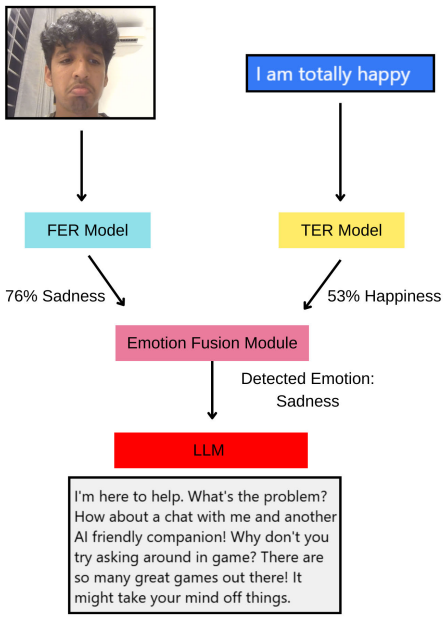


FIGURE 1. Proposed workflow.

Combining ViT-Tiny and MiniLM for real-time systems offers a novel approach for balancing computational efficiency and performance in resource-constrained environments. ViT-Tiny for FER is a lightweight version of the full ViT model optimized for real-time and resource-constrained device deployment [3]. MiniLM, a compact transformer model, excels at natural language processing by offering strong performance with fewer parameters, making it ideal for tasks like question answering and sentence classification [4]. Together, these models enable multimodal systems that can handle both visual and textual data in real-time, reducing latency and computational load.

The integration of multimodal AI significantly enhances performance in real-world, resource-constrained environments, enabling more robust and context-aware decision making. This is particularly beneficial in applications such as autonomous vehicles, where real-time driver monitoring systems can analyze both facial expressions and voice cues to detect fatigue or distraction. Similarly, in smart assistants, leveraging both textual and visual inputs can improve human-computer interactions by adapting responses based on user sentiment. Additionally, edge AI devices, such as wearable health monitors or smart home systems, can process multimodal inputs locally, ensuring low-latency inference and enhanced privacy without relying on cloud-based computations.

Emotion-based chatbots, a form of social robotics, offer significant benefits to society by enhancing human-computer interactions and providing personalized support in various domains. These chatbots make communication more engaging and immersive by recognizing and responding to human emotions and improving user experience and emotional well-being. They promote active interaction and encourage users to express themselves while receiving empathetic responses, which can be particularly beneficial for mental health support, education, and customer service. In educational settings, emotion-aware chatbots can foster critical thinking, problem-solving, and collaboration by adapting to students' emotions and learning needs. Moreover, they hold great potential in assisting individuals with autism or emotional challenges by helping them understand and respond to social cues in a structured and supportive manner. By providing tailored interactions and avoiding sensory overload, these chatbots create a more inclusive environment in which technology fosters emotional intelligence and human connection, ultimately contributing to educational, personal, and professional success [5].

The outputs from both models are then combined using an emotion fusion mechanism, that determines the predominant emotion by comparing the probabilities from both modalities. The selected emotion is used to inform DialoGPT, a language model fine-tuned to generate emotionally aware responses. The resulting system aims to enhance the user experience by adapting the conversational agent's responses to the user's emotional state, thereby promoting more personalized and empathetic interactions [6]. Through a detailed evaluation of the facial and textual emotion recognition models, as well as the fusion mechanism, this study demonstrates how MER can improve the emotional intelligence of virtual assistants.

While unimodal emotion recognition struggles with handling ambiguous or missing data, this study proposes a multimodal fusion mechanism that ensures robustness against noisy inputs. By integrating a lightweight FER model (ViT-Tiny) and TER model (MiniLM), the system achieves real-time performance while maintaining high accuracy.

Unlike conventional virtual assistants, which rely solely on textual inputs, this approach incorporates emotion-aware response generation using fine-tuned DialoGPT, making the AI more adaptive and empathetic. Furthermore, this framework is designed to be computationally efficient, making it deployable in autonomous vehicles, smart assistants, edge AI devices, and social robotics applications.

The major contributions of this study are as follows:

1) Proposal of a multi-modal emotion recognition framework for resource-constrained environments
2) Development of an effective preprocessing pipeline for text-based emotion recognition systems
3) A method that emphasizes extensibility to other large language models (LLMs) and has broader applicability in future multimodal chatbots.

The remainder of this study is organized as follows : Section II examines the relevant literature, Section III describes the proposed methodology, Section IV discusses the results obtained from experimental testing, Section V

concludes the research paper and Section VI suggests future improvements and scope.

II. RELATED WORK

Recent studies have explored emotion recognition, its application in virtual assistants and ways to enhance these systems. This section presents the insights and approaches described in these studies.

A. MULTIMODAL EMOTION RECOGNITION

Recent advancements in MER leverage the integration of multiple data types, such as speech, visual, and text, which leads to a higher accuracy in the classification of emotions through deep learning approaches. Key research indicate that deep learning models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Bi-directional Long Short-Term Memory networks (BiLSTMs), Generative Adversarial Networks (GANs), and transformers, play a key role in advancing MER, owing to their ability to process high-dimensional complex data, and drawing out intricate patterns [7], [8]. Traditional machine learning algorithms, including Support Vector Machines (SVMs), are also utilized in MER, although their performance tends to lag in capturing the nonlinear patterns inherent to multivariate time-series data, particularly when datasets are limited [8].

Attention-based fusion strategies, as highlighted by Geetha et al. [7], have emerged as promising approaches for integrating features from various modalities for emotion recognition. By assigning weights to prominent features, these methods improve the representation of emotional states while addressing the challenges posed by the heterogeneity of multimodal data. However, high-dimensional features often introduce integration difficulties and processing inefficiencies, leading to trade-offs between the performance and computational constraints. Similarly, Tzirakis et al. [9] demonstrated the efficacy of multimodal fusion by integrating speech and visual data using CNNs for speech signals and ResNet50 for visual inputs. Using the RECOLA dataset and eGeMAPS acoustic parameter set, this study demonstrates how integrated multimodal cues can provide richer representations of emotional states, making models more adept at understanding complex affective cues.

Ahmed et al. [8] highlighted the advantages of multimodal datasets such as emoFBVP, DEAP, SEED, FER2013, K-EmoCon, PMEmo, MAHNOB- HCI, VREED, and MEmoR in advancing MER research. These datasets provide standardized and diverse inputs that facilitate the generalizability of CNNs, RNNs, and Autoencoders, whose hierarchical architectures excel in high-level feature extraction. Despite these advances, limitations such as insufficient sample sizes and challenges with high-dimensional feature computation hinder the real-world application performance. Pan et al. [10] reiterated these concerns, pointing out that data set variability and scale significantly constrain model generalization across domains. Although hybrid fusion strategies, which combine feature-level and decision-level fusion, improve accuracy, they introduce greater complexity and computational demands. Moreover, issues such as subjectivity in emotion annotation and noise in physiological signals call for advanced preprocessing techniques to ensure data reliability and model robustness.

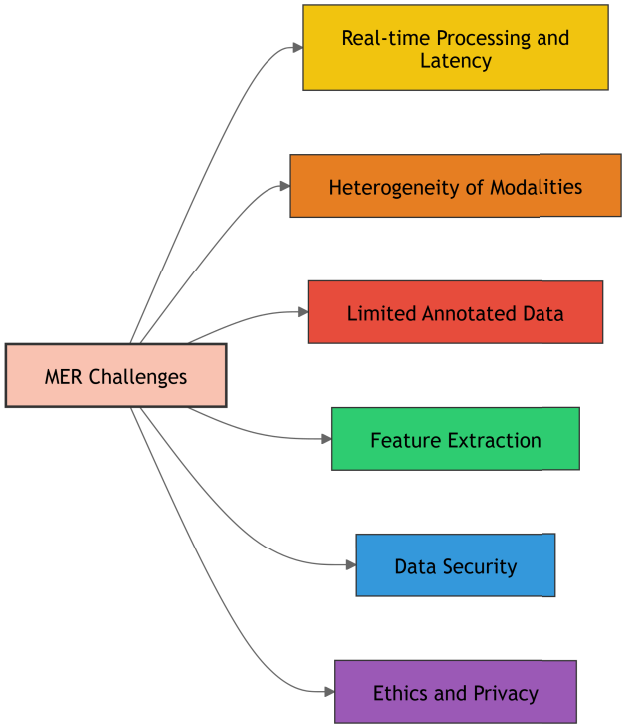


FIGURE 2. Challenges related to MER.

In summary, recent literature points to the significant role of deep learning in MER, highlighting the success of various architectures and fusion strategies in addressing the heterogeneity of multimodal data. However, computational and data set limitations remain, leaving room for future work to further optimize these models for practical and efficient emotion recognition applications(Fig 2).

B. TEXTUAL EMOTION RECOGNITION

Research on TER has brought forward both innovative methods and ongoing challenges.

Deng and Ren [1] and Machová et al. [11] highlighted key limitations of TER, with an emphasis on the challenges posed by imbalanced datasets and ambiguous emotional boundaries. Deng et al. reviewed the inconsistency in emotional annotation schemes across datasets such as SemEval2007, ISEAR, EEC and EMOBANK, as well as imbalanced distributions across emotion categories which complicates effective model training and evaluation. It was noted that textual expressions showcase fuzzy emotion boundaries, resulting in multiple emotion labels, complicating classification tasks owing to ambiguity. Similarly, Machová et al. pointed out that chatbots face difficulty in understanding the full context of user inputs, especially when training data is

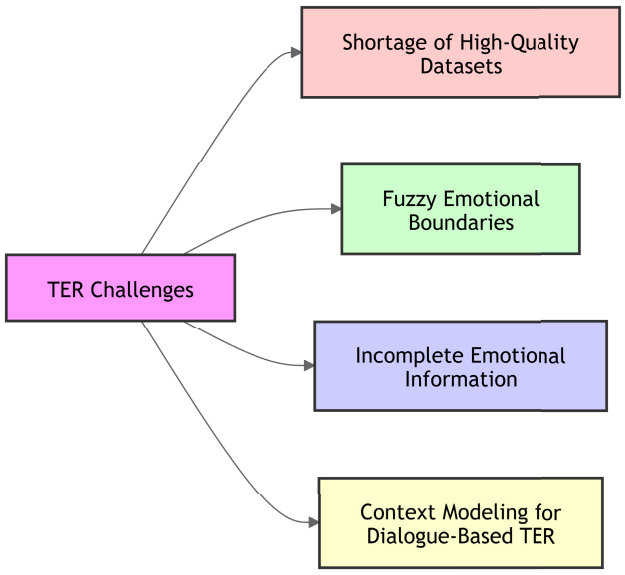


FIGURE 3. Challenges related to TER.

limited or language is ambiguous. Both studies advocated multimodal approaches, integrating text with other modalities like facial expressions and vocal cues to provide richer emotional context and enhance accuracy. Machová et al. additionally proposed leveraging deep learning models with attention mechanisms to address complex expressions such as sarcasm and to improve TER's contextual understanding.

Poria et al. [12] and Seyeditabari et al. [13] delved deeper into the semantic intricacies and interpretive challenges of TER. Poria et al. illustrated how emotional perception varies depending on the individual's background and context, exemplified by the statement "Lehman Brothers' stock is plummeting!!," which could signify distress or excitement depending on the speaker's perspective. Seyeditabari et al. further explore dthese challenges through idiomatic expressions, demonstrating how phrases like "he lost his cool" or "you make my blood boil" often fail to be accurately classified due to their non-literal meanings. Both reviews underscore the limitations of text-based emotion recognition and advocate for advancements in contextual understanding and refinement of deep learning models. Incorporating multimodal inputs, such as vocal tone and facial expressions, along with improved model interpretability, could address these challenges and pave the way for more reliable emotion detection systems.

Further advancements specific to sentiment analysis have been explored by Phan et al. [14] who proposed FeDN2, a sentence-level sentiment analysis model that integrates fuzzy logic into a deep neural network. It combines BERT embeddings, a fuzzification layer using Gaussian membership functions, and deep CNN blocks, followed by defuzzification and classification. This architecture improved the handling of ambiguous sentiment expressions. Maleszka [15] surveyed explainable sentiment analysis (XSA) techniques, including attention-based visualization, and post-hoc methods like LIME and SHAP. The survey emphasized the need for interpretability in sentiment models to enhance user trust and model transparency.

Together, these studies underscore the limitations of TER when relying solely on text and highlight the potential of multimodal approaches to enrich emotion recognition by capturing a broader range of emotional signals. By incorporating supplementary modalities and refining deep learning models, these approaches can help overcome challenges like imbalanced data, fuzzy emotional boundaries, and privacy concerns, making emotion recognition systems more accurate, empathetic, and secure (Fig 3).

C. FACIAL EMOTION RECOGNITION

Recent advancements in FER highlight both the potential and limitations of current deep learning models and datasets for achieving high accuracy across various applications. Harika et al. [16] explored the performance of various datasets in the Facial Emotion Recognition (FER) domain, highlighting a trade-off between the dataset size and robustness. While smaller datasets such as CK+, JAFFE, KDEF, and FERG enable models to achieve impressive accuracies of up to 99.8%, their limited diversity constrains their applicability in real-world scenarios. In contrast, larger datasets such as FER-2013, EMOTIC, and CAER-S offer greater diversity and volume, enhancing model generalizability and reliability in complex, real-world environments. Complementing this analysis, Chowdary et al. [2] investigated the role of transfer learning in FER, emphasizing its utility in leveraging pre-trained networks such as ResNet50, VGG19, Inception V3, and MobileNet for improved accuracy. Their study underscores how deep learning eliminates the need for separate feature extraction and classification phases, streamlining the process and achieving high performance. For instance, MobileNet demonstrated superior results with an F1 score of 0.93 and an accuracy of 98%, attributed to its efficiency and compact architecture.

Both studies highlighted the importance of advancing FER techniques using larger datasets and more efficient models. Harika et al. stressed the need for datasets with greater diversity to address real-world challenges, whereas Kalpana et al. emphasized data augmentation and transfer learning as critical tools for optimizing model performance. Together, these works suggest that integrating FER with additional modalities such as speech and EEG signals, could further enhance model capabilities and application scope. By addressing dataset limitations and refining learning strategies, these advancements pave the way for the development of robust, real-time FER systems suitable for human-computer interaction and other real-world applications.

D. EMOTIONALLY AWARE VIRTUAL ASSISTANTS

Recent research on Emotion Detection in Virtual Assistants and chatbots emphasizes the potential of emotionally

intelligent systems to enhance user interactions across various domains, from customer service to healthcare.

Mishra [17] and Bilquise et al. [18] emphasized the potential of emotionally intelligent systems to enhance user interactions in domains ranging from customer service to healthcare. Mishra et al. explored the capabilities of advanced transformer models to generate contextual responses, enabling virtual assistants to hold coherent and meaningful conversations. By recognizing and responding to users' emotional cues, these virtual assistants can provide empathetic experiences, that is valuable in applications such as mental health support and personal assistance. However, they identified challenges such as the diversity and quality of training data, subjectivity in evaluating responses, and the critical need for data privacy. They proposed integrating long-term memory and user context to deliver personalized, contextually aware interactions that adapt to user preferences and past behaviors.

In their review, Bilquise et al. similarly underscored the importance of emotion recognition in improving chatbots. They examined the use of lexicon-based approaches, particularly the Valence, Arousal, and Dominance (VAD) vector, to enable fine-grained emotion detection in text-based interactions. While Seq2Seq models are commonly used for response generation, Bilquise et al. noted that these models often produce repetitive or ''dull'' responses, motivating a shift toward RNNs and LSTMs to enhance conversational quality. One of the main challenges identified is the lack of emotional labeling in open-domain conversational datasets, which requires significant preprocessing through classifiers, lexicon-based, or hybrid methods. They advocate for the development of task-oriented, emotionally intelligent chatbots, suggesting that domain-specific, voice-based, and multimodal chatbots could achieve more sophisticated, context-sensitive interactions [18].

Together, these studies highlight the promise and challenges of creating emotionally aware VAs and chatbots. Both Mishra et al. and Bilquise et al. stressed that emotionally intelligent systems can greatly enhance user experience by responding to emotional cues and adjusting to user contexts. However, challenges such as data diversity, privacy concerns, and limitations in existing conversational datasets point to the need for more robust, multimodal approaches (Fig 4). Future work in this area will likely focus on integrating long-term user context, refining emotion detection across multiple modalities, and creating domain-specific solutions that can handle complex emotional nuances, ultimately leading to more accurate, personalized, and secure interactions [17], [18].

E. RESPONSE GENERATION IN VIRTUAL ASSISTANTS

Recent advancements in the development of empathetic chatbots have leveraged Natural Language Processing (NLP) techniques to improve both emotion detection and empathetic response generation. Abuhmida et al. [19] developed a hybrid model that combines emotion detection and empathetic response generation using BiLSTM and Linear SVC for emotion categorization, achieving high accuracy in recognizing sequential data. They fine-tuned DialoGPT and a transformer-based model on datasets like EmpatheticDialogue and DailyDialogue, with the transformer outperforming DialoGPT in perplexity, though challenges like emotion label imbalance and high computational demands remained. Similarly, Wang et al. [20] enhanced DialoGPT to create ''Emily,'' a chatbot fine-tuned on Twitter and empathetic dialogue datasets using emoji-based sentiment labels. Using Maximum Mutual Information (MMI) scoring, Emily minimized generic responses and guides users from negative emotions to positivity. Wang et al. also incorporated a RoBERTa-based classifier and Knowledge Graph Question Answering (KGQA) module to ensure persona consistency and provide knowledge-based, coherent responses, improving engagement and emotional sensitivity. Both studies highlight the potential of advanced algorithms and fine-tuned models in building emotionally aware, empathetic chatbots despite challenges such as data imbalance and computational constraints.

Booth et al. and Wang et al. highlighted the need for better emotion detection algorithms and more diverse datasets to improve performance on subtle emotions. Future directions include multimodal data integration (e.g., voice and facial cues) to bolster emotion recognition, enhancing chatbots' robustness in various real-world applications such as customer support and mental health. These efforts indicate that, with targeted model training and persona integration, DialoGPT-based chatbots are well-positioned to support

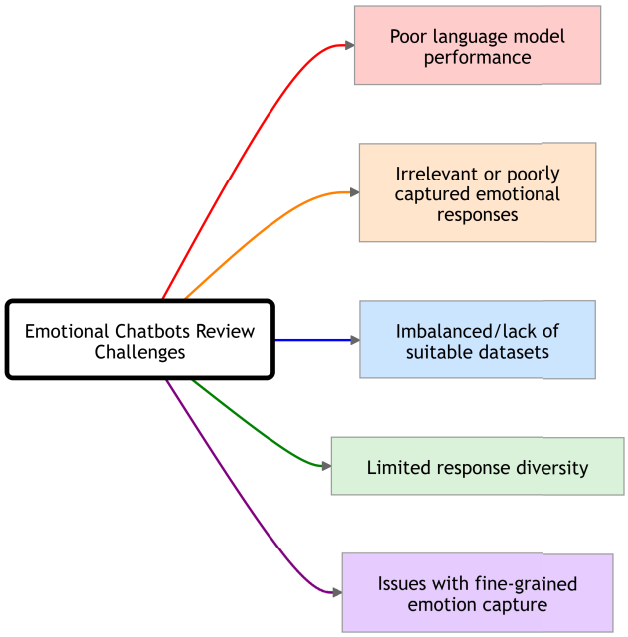


FIGURE 4. Existing emotional chatbot challenges.

emotionally aware, empathetic interactions in a range of contexts.

F. EXAMPLES OF VIRTUAL ASSISTANTS INCORPORATING MULTIPLE MODALITIES

Recent advancements in conversational agents for virtual environments and smart systems have demonstrated the potential of multimodal and emotionally aware designs to enhance user engagement and response accuracy. The metabot "Demic," [21], and the emotionally aware campus assistant presented by Chiu et al. [22] demonstrate how multimodal designs can improve user interaction in virtual environments. Demic leverages speech and text inputs alongside a dialogue register that stores prior interactions, enabling continuity and coherence in extended dialogues. Neural network-based classification selects appropriate responses, improving the relevance of replies, whereas emotion recognition focuses on identifying specific negative emotions to enhance sensitivity. Similarly, Chiu et al. integrated speech recognition, emotion detection, and augmented reality (AR) to create a personalized and immersive smart campus assistant. Utilizing deep neural networks (DNNs), Word2Vec embeddings, and RNN/LSTM models, their system captures linguistic and temporal dependencies effectively, achieving an impressive 95.6% accuracy in emotion recognition for short sentences.

Other studies include Framework for Improving Learning Through Webcams And Microphones (FILTWAM) proposed by Bahreini et. al. [23] with the goal of validating the use of webcam data for a real-time and adequate interpretation of facial expressions into extracted emotional states and Duwenbeck et. al. which proposed Project "AudEeKA", whose aim is to use speech and other bio signals for emotion recognition to improve remote, but also direct, healthcare [24].

Together, these studies underscore the importance of integrating multimodal capabilities and memory-based context preservation to improve the interaction quality in virtual environments and intelligent systems. Both Demic and the smart campus assistant leverage neural networks for contextual response generation and employ sophisticated emotion recognition techniques to respond more empathetically to user input. Future developments in these areas are likely to focus on expanding emotional sensitivity across a broader range of emotions, refining spoken language models, and further enhancing real-time response mechanisms. These improvements could enable more nuanced, context-aware, and emotionally intelligent agents across applications from social virtual worlds to smart campus environments.

The challenges outlined in Section II highlight the limitations of unimodal approaches and the need for a robust multimodal emotion recognition system. This study directly addresses these challenges by integrating FER and TER using ViT-Tiny and MiniLM, respectively. The proposed emotion fusion mechanism ensures adaptability to ambiguous or conflicting cues, thereby improving accuracy. By incorporating lightweight architectures, this approach also addresses computational efficiency concerns, making it deployable in real-time applications.

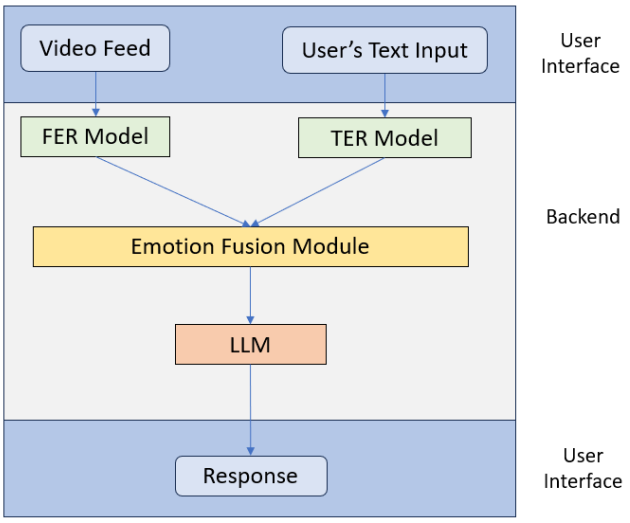


FIGURE 5. Block diagram of proposed architecture for virtual assistant.

While prior works have explored multimodal emotion recognition and empathetic dialog generation, few have addressed the practical constraints of real-time deployment. Moreover, integrating TER-FER pipelines directly with pre-trained transformers like DialoGPT for emotion-conditioned response generation remains underexplored. Our study bridges this gap by introducing a seamless FER-TER-LLM flow suitable for real-time empathetic virtual assistants.

III. PROPOSED METHODOLOGY

The proposed methodology aims to enhance conversational agents by integrating multimodal emotion recognition from visual and textual data. A video feed is analyzed by an FER model based on ViT-Tiny, whereas the user's textual input is processed by a TER model based on MiniLM. The outputs from these models are combined using an emotion fusion mechanism that compares the probability distributions of detected emotions from both modalities and selects the most prominent emotion. This selected emotion is then used as input to a language model (DialoGPT) to generate emotionally aware responses. The system architecture enables a more nuanced interaction by allowing the conversational agent to respond based on the user's detected emotional state, leading to more personalized and empathetic communication (Fig 5). The proposed methodology is described in the form of an algorithm (Algorithm 1) as follows :

- **Step 1: Initialization** – Load pre-trained models for facial emotion recognition (FER) using ViT-Tiny, textual emotion recognition (TER) using MiniLM, and response generation using DialoGPT.

Algorithm 1 Multimodal Emotion Recognition and Response Generation

Input: Facial Image I, Textual Input T
Output: Emotionally-Aware Response R
Step 1: Initialize Models
Load Pre-trained ViT-Tiny for FER (M_{FER})
Load Pre-trained MiniLM for TER (M_{TER})
Load Fine-tuned DialoGPT (M_{Dialog})
Step 2: Preprocessing
$I_{pre} \leftarrow Preprocess(I)$ ▷ Resize, Normalize Image
$T_{pre} \leftarrow Preprocess(T)$ ▷ Clean and Tokenize Text
Step 3: Facial Emotion Recognition (FER)
$P_{FER} \leftarrow M_{FER}(I_{pre})$ ▷ Get Emotion Probabilities from Image
Step 4: Textual Emotion Recognition (TER)
$P_{TER} \leftarrow M_{TER}(T_{pre})$ ▷ Get Emotion Probabilities from Text
Step 5: Emotion Fusion Mechanism
for $e \in Emotion\_Classes$ **do**
    $P_{final}(e) \leftarrow \max(P_{FER}(e), P_{TER}(e))$
end for
$Emotion_{final} \leftarrow \arg\max(P_{final})$
Step 6: Generate Emotionally-Aware Response
$R \leftarrow M_{Dialog}(T, Emotion_{final})$ ▷ Generate Emotion-Aware Response
Step 7: Output Response
Display Response R to User
End Algorithm

- **Step 2: Preprocessing** – Capture both facial and textual inputs in real-time. Preprocess the image and text inputs to prepare them for analysis by their respective models.
- **Step 3: FER** – Use ViT-Tiny to classify emotions from preprocessed facial images.
- **Step 4: TER** – Use MiniLM to classify emotions from preprocessed text inputs.
- **Step 5: Fusion Mechanism** – Combine the emotion probabilities from both modalities and select the dominant emotion using the maximum probability across all classes.
- **Step 6: Response Generation** – Generate an emotion-aware response using DialoGPT based on the detected emotion and the contextual input text.
- **Step 7: Output** – Display the final response to the user.

A. FACIAL EMOTION RECOGNITION

The objective of this study was to identify an optimal deep learning model for real-time FER in human–computer interaction applications.

1) DATASETS USED AND DATA PREPARATION

Three widely used datasets were examined in this study: CK+ [25], FER2013 [26], and AffectNet [27].These datasets vary in classes and nature, hence requiring different preprocessing techniques to prepare the data for use in deep learning models.

The CK+ dataset is a widely recognized facial expression dataset that contains 593 sequences from 123 subjects. Each sequence consists of a series of images showing gradual facial expression changes, typically in response to an emotional trigger. It covers seven emotion classes: anger, contempt, disgust, fear, happiness, sadness, and surprise [25]. The images from the CK+ Dataset were read as grayscale and resized to fit the input parameters of the deep learning models. The images and labels were converted to numpy arrays for efficient processing. Label encoding was performed to obtain numerical labels for model training.

The FER2013 dataset, which is available as part of Kaggle's facial emotion recognition challenge, consists of 35,887 labeled images depicting various facial expressions across seven categories: angry, disgust, fear, happy, sad, surprise, and neutral. The images are 48×48 pixels in size, and the dataset is particularly challenging due to its noisy data and imbalanced class distribution [26]. To work with this imbalance of data, the Focal Loss Function was used for its ability to focus on hard-to-classify items.

$$\text{Focal Loss} = -\alpha \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$$

where:

- p_t is the predicted probability of the true class.
- α is a scaling factor used to balance the importance of different classes.
- γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted.

Landmark-based cropping was performed using the DLIB library, which extracted 68 facial landmarks. DLIB's face detector, leveraging Histogram of Oriented Gradients (HOG) features and a linear classifier, identified key facial regions such as the eyes, eyebrows, and lips - features considered most indicative of emotional states [28]. These regions of interest (ROI) were cropped, resized, and converted to NumPy arrays. Label encoding was performed to obtain numerical labels for model training.

AffectNet is a comprehensive facial expression dataset comprising approximately 400,000 images (collected from the web) that have been manually annotated for eight different facial expressions: neutral, happy, angry, sad, fear, surprise, disgust, and contempt [27].

For preparing the data, TensorFlow's ImageDataGenerator was employed to augment the training dataset. Several transformations, including rotation, shifting, shearing, zooming, and horizontal flipping, were applied to improve model robustness and generalization. The images were then resized to match the input size requirements of the deep learning models. Pixel values were also rescaled to the range (0, 1).

2) MACHINE LEARNING MODELS

For the emotion recognition tasks, several pre-trained deep learning models were explored, including ResNet50,

MobileNet, and EfficientNet along with ViT and its model checkpoints. The specific configurations of these models were adjusted to optimize performance on each dataset.

ResNet50, a deep convolutional neural network with 50 layers, is built on the concept of residual connections that bypass layers to address the vanishing gradient problem and enable effective training of deep networks. Its architecture uses bottleneck blocks with 1×1 and 3×3 convolutions, followed by global average pooling and a fully connected layer for classification. It is well-suited for emotion recognition tasks due to its ability to handle complex features, achieve high accuracy, and perform effectively on large-scale datasets [29].

EfficientNet is a family of models designed for accuracy with computational efficiency. Its architecture features MBConv blocks, which combine depthwise separable convolutions with inverted residuals for improved performance. EfficientNet achieves high accuracy with fewer parameters, making it ideal for resource-constrained tasks. The model's scalability (EfficientNet-B0 to B7) and efficiency make it a strong choice for emotion recognition applications [30].

MobileNet is a lightweight deep learning model family optimized for mobile and edge devices, utilizing depthwise separable convolutions for efficiency, offering flexible scaling through width multipliers, and typically trained on ImageNet for effective image classification and recognition tasks. MobileNet's small model size, fast inference, and adaptability through the width multiplier make it ideal for real-time emotion recognition on resource-constrained devices [31].

ViT is a deep learning model that uses transformer-based architectures, originally designed for natural language processing, to process images. Unlike traditional convolutional neural networks (CNNs), ViT divides an image into patches and processes them in parallel, allowing it to capture long-range dependencies and subtle patterns in the image [3]. ViT-Tiny, a smaller version, balances performance and efficiency, making it ideal for resource-constrained devices. Compared to MobileNet, ViT-Tiny is better at capturing subtle facial features, making it more effective for facial emotion recognition while still being efficient for real-time deployment (Fig 6).

3) TRAINING AND TESTING THE MODELS

The chosen deep learning models, were evaluated across multiple datasets; CK+ [25], FER2013 [26], and AffectNet [27], with variations in data balancing, loss functions, and architecture modifications. Each model was trained with early stopping constraints and restoring of best weights to ensure optimal performance for each variation. The performance of each configuration was analyzed based on training and test accuracy, with an emphasis on real-time performance when integrated with OpenCV's Haarcascade face detection classifier (Table 1).

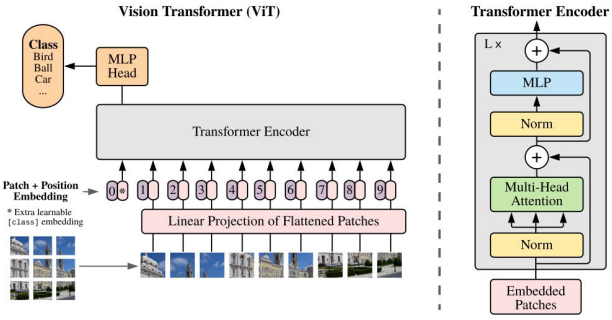


FIGURE 6. ViT architecture.

4) OBSERVATIONS

While CK+ yielded high accuracy of 100% during both training and testing, it failed in real-time applications. The model consistently classified faces into the same emotion category, indicating overfitting and poor generalization. This dataset's small size and limited diversity restricted the model's ability to perform well in practical scenarios. On FER2013, MobileNet achieved a test accuracy of 51%, but real-time performance was unsatisfactory. The limited diversity in emotion categories and insufficient generalization in FER2013 led to inaccurate classifications during live testing. Even though data balancing methods, such as Near Miss undersampling and SMOTE, were applied, FER2013 lacked the emotional nuances needed for robust emotion recognition.

Understanding the limitations of FER2013 and CK+, it was decided that a balanced subset of the AffectNet dataset, with its larger size and more diverse emotional categories, would be a better dataset for testing. While AffectNet offers greater diversity and includes images captured under challenging conditions—such as poor lighting, non-frontal poses, and partial facial occlusions—these difficult cases constitute a smaller portion of the dataset. As a result, they are underrepresented during training, which may limit the model's ability to generalize effectively in real-world scenarios where such conditions are more prevalent.

EfficientNet showed promise as a high-performing CNN model, however it was recognized that transformer-based models, in general, offer superior performance in capturing complex patterns, leading to the selection of one of the variants of the ViT model [3].

5) ADOPTED FER MODEL ARCHITECTURE FOR VIRTUAL ASSISTANT

To adapt the model to the task of facial emotion recognition, transfer learning is applied using the AffectNet dataset. AffectNet's extensive variety of expressions and demographic diversity make it an essential tool for advancing research in facial expression analysis. Amongst the transformer-based ViT models, ViT-Tiny achieved a test accuracy of $70.69 \pm 1.03\%$ on AffectNet(Table 1). While

TABLE 1. Comparison of various model architectures and configurations for emotion recognition.

MODEL	DATASET	DLIB FILTERED	LOSS HYPERPARAMETER	OPTIMIZER	Data Balancing	ARCHITECTURE MODIFICATIONS	TRAINING ACCURACY	TEST ACCURACY
MOBILENET	FER2013	NO	FOCAL LOSS: gamma = 2; alpha = 0.25	adam	NO	NO	45.19%	48%
	FER2013 w/o disgust	YES					94%	51%
				rmsprop	YES: Near Miss Undersampling & SMOTE		31.42%	16%
				adam			98.53%	51%
					NO	Added Dropout Layers and LR Regression	42.52%	41%
					YES: Near Miss Undersampling & SMOTE		26.36%	29%
	CK+	NO	categorical_crossentropy	adam (lr = 0.0001)	NO	NO	100%	100%
RESNET	AffectNet					NO	93.84%	66.73%
							98.4%	68.55%
EFFICIENTNET							96.54%	70.24%
	CK+			adam		NO	97.95%	99%
					YES: Near Miss Undersampling & SMOTE	Dropout layer before output layer	83.86%	93%
	FER2013 w/o disgust	YES	FOCAL LOSS: gamma = 2; alpha = 0.25		NO	NO	65.62%	57%
ViT-Base	AffectNet	NO	categorical_crossentropy	adam	YES	NO	98.49%	68.17%
ViT-Small							97.76%	71.39%
ViT-Tiny							**97.94**%	**70.69**%

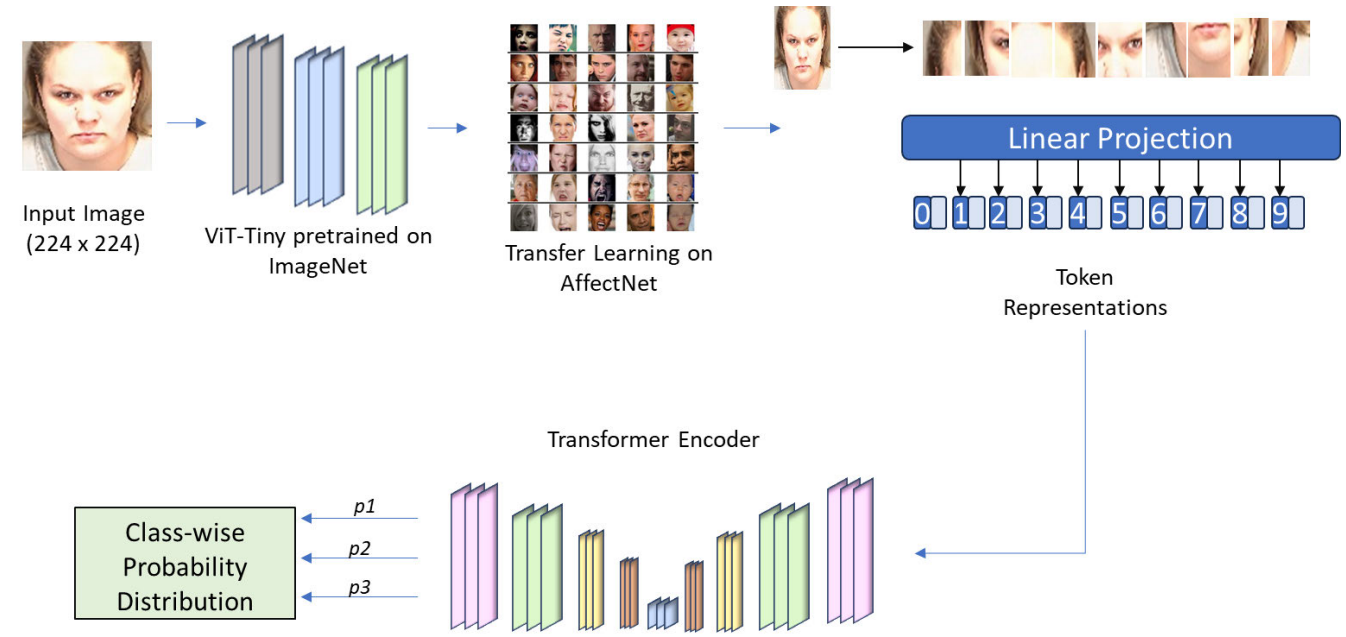


FIGURE 7. Adopted architecture for facial emotion recognition.

ViT-Small also yielded comparable accuracy, ViT-Tiny was selected due to its lightweight nature. Considering other lightweight models such as EfficientNet (70%), ResNet (68%), and MobileNet (67%), ViT-Tiny performs on par with these and offers a balanced trade-off between performance and model complexity, making it a favorable choice for resource-constrained settings. The adopted architecture is illustrated in Fig 7.

The trained model is integrated with OpenCV's Haar-cascade classifier for efficient face detection. The Haarcascade classifier, a lightweight and real-time face detection algorithm, is used to locate faces in a video stream or image input [32]. Once a face is detected, it is cropped and resized to 224×224 pixels to match the input size requirements of the ViT-Tiny model. This preprocessing step ensures that only the relevant facial regions are fed into the deep learning model, enhancing the accuracy of emotion classification.

B. TEXTUAL EMOTION RECOGNITION

For TER, this study makes use of the CARER [33] and GoEmotions [34] datasets. GoEmotions, which has 58,000 entries and 27 fine-grained classes, was mapped to Ekman's six basic emotions plus neutral for consistency with facial emotion recognition, whereas CARER offers 416,809 entries across six primary emotions.

To ensure clean input for models, a strong preprocessing pipeline was used, which included noise removal, case normalization, contraction expansion, and spelling correction. CARER experiments revealed that Logistic Regression performed best with Count Vectorizer (88% accuracy). MiniLM, a lightweight transformer, was chosen for its effectiveness and suitability for real-time applications after achieving 59% accuracy on GoEmotions. For scalable TER, this methodology combines preprocessing, effective models, and high-quality datasets.

1) DATASETS USED

The textual emotion recognition experiments involved exploration of various models on two different datasets - CARER and GoEmotions.

The CARER (Contextualized Affect Representations for Emotion Recognition) [33] dataset is a widely used resource for emotion recognition in textual data. It contains 416,809 entries categorized into six distinct emotional classes: sadness, joy, love, anger, fear, and surprise.

The GoEmotions dataset, developed by Google Research, was selected for its fine-grained emotional labels and its suitability for multi-class classification tasks. It is considered one of the largest fully annotated English-language fine-grained emotion datasets available, making it particularly valuable for training deep learning models capable of distinguishing subtle emotional cues. The dataset also offers a diverse set of emotional expressions, which is ideal for both text-based and multimodal emotion recognition tasks. It consists of 58,000 entries annotated with 27 distinct emotion classes [34]. For the purposes of this study, the emotion classes were grouped to fit into Ekman's six basic emotions(joy, sadness, anger, surprise, neutral, disgust, fear), plus a neutral category. The grouping was done in the following manner

- Joy: "admiration", "amusement", "approval", "excitement", "gratitude", "joy", "love", "pride", "optimism"
- Sadness: "disappointment", "grief", "remorse", "sadness"
- Anger: "anger", "annoyance", "disapproval"
- Surprise: "surprise","realization"
- Neutral: "neutral", "caring", "curiosity", "confusion", "desire", "nervousness".
- Disgust: "disgust", "embarrassment".
- Fear: "fear"

2) TEXT PREPROCESSING PIPELINE

In this study, we implemented a comprehensive text preprocessing pipeline designed to clean and normalize textual data. The preprocessing pipeline consists of the following stages (Fig 8) :

a: BASIC TEXT CLEANING

The clean_text function removes noise and irrelevant elements from the raw text data:

- Emoji Demojization: Converts emojis into textual representations using the emoji library. Emoji demojization and subsequent removal simplify textual inputs for analysis [35], [36].
- Case Normalization: Converts all text to lowercase for uniformity. Lowercasing is a common preprocessing step to reduce dimensionality while preserving semantic meaning [37].
- Removal of Unnecessary Tokens: Regular expressions are employed to remove square bracketed content, HTML tags, URLs, newline characters, and alphanumeric tokens, reducing noise and preserving meaningful text [38], [39].

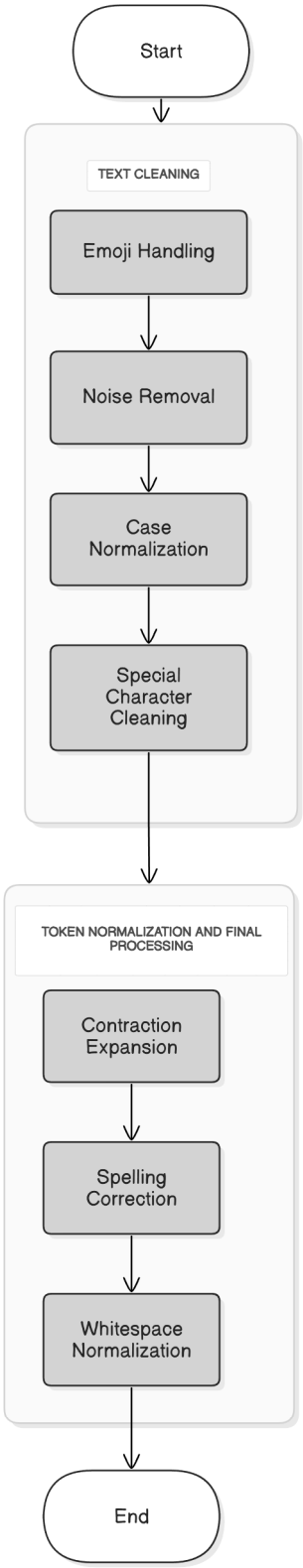


FIGURE 8. Text preprocessing flow.

TABLE 2. Comparison of classifiers using different vectorizers on CARER dataset.

Vectorizer	Classifier	Accuracy	Weighted Recall	Weighted avg F1 Score
TFIDF	Logistic Regression	0.85375	0.85	0.85
	SVM	0.84896	0.85	0.84
	Gradient Boosting	0.84042	0.84	0.84
	Random Forest	0.81042	0.81	0.8
	Multinomial Naïve Bayes	0.57083	0.57	0.47
Count Vectorizer	Logistic Regression	0.88188	0.88	0.88
	SVM	0.87750	0.88	0.88
	Random Forest	0.84208	0.84	0.84
	Gradient Boosting	0.83729	0.84	0.84
	Multinomial Naïve Bayes	0.65500	0.66	0.59

- Character Restriction: Retains only alphabetic characters and punctuation, ensuring a focus on linguistically meaningful content [40], [41].

b: CONTRACTION EXPANSION

The clean_contractions function expands contractions based on a predefined mapping dictionary. This step resolves ambiguities caused by contractions and ensures text consistency, aligning with prior studies emphasizing normalization for better model understanding [42].

c: SPECIAL CHARACTER CLEANING

The clean_special_chars function handles:

- Custom Mapping: Replaces specific special characters with standardized forms based on a mapping dictionary, ensuring text uniformity.
- Punctuation Handling: Provides consistent spacing around punctuation marks to improve tokenization accuracy.
- Unicode and Language-Specific Cleaning: Handles problematic Unicode characters (e.g., zero-width spaces) and domain-specific tokens, ensuring a cleaner text corpus.

d: SPELLING CORRECTION

The correct_spelling function uses a dictionary of common misspellings to standardize variations. Previous work highlights the importance of spelling correction in enhancing semantic coherence and improving downstream natural language processing (NLP) tasks.

e: WHITESPACE REMOVAL

The remove_space function eliminates extra spaces and ensures proper tokenization. Space normalization avoids redundant padding during text vectorization, leading to better feature extraction.

f: PIPELINE INTEGRATION

The text_preprocessing_pipeline combines the above functions, applying them sequentially to produce clean, standardized text. The pipeline can be applied to datasets programmatically for large-scale preprocessing.

TABLE 3. RNN-based model performances on the Go Emotions dataset.

Model	Validation Accuracy
GloVe Embeddings + Bi-Directional LSTM	0.5746
LSTM + Attention Layer	0.5272

TABLE 4. Transformer-based model performances on the go emotions dataset.

Model	Eval Accuracy	Eval F1
BERT-Base	0.5716	0.5643
RobertA	0.4213	0.2497
DistilRobertA	0.5725	0.5505
DeBERTA	0.5874	0.5776
MiniLM	0.5922	0.5743

3) TRAINING AND TESTING MODELS

Experiments were initially done on the CARER dataset using two different vectorization techniques - TFIDF and Count Vectorizer. After vectorization i.e conversion of textual data into numerical format, classical machine learning models were tested out including Logistic Regression, SVM, Gradient Boosting, Random Forest and Multinomial Naïve Bayes. Count Vectorizer with Logistic Regression proved to be the best performing model with an accuracy of 88% on this dataset.

For emotion classification on the GoEmotions Dataset, three different types of models are explored - RNN based models including BiLSTM and LSTM, transformer based models including MiniLM, DistilRoberta, Roberta and DeBert and finally the best performing model on the CARER dataset, Count Vectorizer with Logistic Regression (Table 3, Table 4). Count Vectorizer + Logistic Regression yielded an accuracy of 57%, similar to most models in the tables. MiniLM proved to be the model with better performance with an accuracy of 59%.

4) ADOPTED TER MODEL ARCHITECTURE FOR VIRTUAL ASSISTANT

Out of all the models tested, MiniLM on the GoEmotions dataset was chosen due to its higher accuracy and due to a class mapping that would match the chosen framework for the FER model. MiniLM is a smaller and more efficient variant of BERT (Bidirectional Encoder Representations from Transformers), designed to be computationally less

expensive while still retaining much of the performance of larger models. MiniLM is particularly well-suited for resource-constrained environments, making it an ideal choice for real-time emotion recognition applications. MiniLM uses masked language modeling and next sentence prediction tasks during pre-training, allowing it to learn rich contextual representations of text. Its architecture enables efficient fine-tuning for downstream tasks such as emotion classification [4].

To supplement the context-aware emotion recognition pipeline, a pre-trained T5-based sarcasm detection model [43] is employed in parallel with the MiniLM-based emotion classifier. Upon detecting sarcasm, a custom rule-based negation mapping is applied to reinterpret the literal emotion prediction. This mapping heuristically transforms the surface-level emotional expression into its likely intended affective counterpart (e.g., sarcastic joy → anger), thus enabling more semantically robust emotion classification in the presence of ironic or sarcastic language.

C. EMOTION FUSION MECHANISM

The emotion fusion mechanism integrates outputs from multiple modalities, such as facial expressions and textual cues, to determine the most dominant emotion. By operating at the decision level, it compares probabilities across modalities, resolving conflicts and ensuring accurate, context-aware emotion detection.

1) CONVENTIONAL METHODS

The conventional multimodal information fusion methods contain feature-level fusion, decision-level fusion, as well as model-level fusion [10].

Feature-level fusion combines features from different modalities into a single vector and feeds it into a classifier for emotion classification. However it can face dimensionality issues when concatenating large feature vectors. It also fails to capture associations across different modalities.

Decision-level fusion combines the results of individual modality classifiers using algebraic rules. This also does not consider intermodal correlations. However, each modality can use the classifier best suited for its specific task.

Model-level Fusion, models each modality individually while considering the correlations between them, allowing for better integration of modality-specific information. It considers intermodal correlations but is more complex and computationally intensive.

Hybrid-level Fusion combines multiple fusion strategies (feature-level, decision-level, model-level) to leverage the strengths of each.

2) ADOPTED EMOTION FUSION MECHANISM

For the Virtual Assistant, this study chose to go forward with a decision-level fusion as explained below.

The emotion fusion mechanism compares the probabilities from both sources and selects the emotion with the highest value from either source. This is done for each emotion category, and the one with the maximum probability is chosen as the overall emotion (Fig 9).

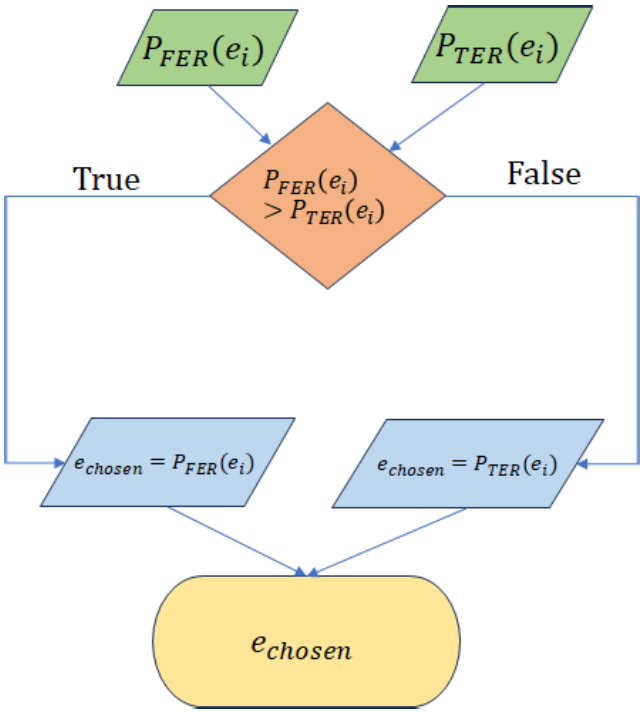


FIGURE 9. Emotion Fusion Mechanism.

Formally, we can define the fusion mechanism as:

$$e_{\text{chosen}} = \arg\max_i \left(P_{FER}(e_i) \cup P_{TER}(e_i)\right)$$

where:

- $P_{FER}(e_i)$ is the probability of emotion e_i from facial recognition,
- $P_{TER}(e_i)$ is the probability of emotion e_i from textual recognition,
- The operator $\cup$ denotes the element-wise maximum, selecting the higher of the two values for each emotion e_i,
- e_{chosen} is the selected emotion based on the highest probability.

An example of the working of the fusion mechanism can be demonstrated as follows.

Let's assume the following probability distributions for a set of emotions (e_1: happy, e_2: sad, e_3: angry):

$$P_{\text{FER}}(e_1) = 0.6, \quad P_{\text{FER}}(e_2) = 0.3, \quad P_{\text{FER}}(e_3) = 0.2$$
$$P_{\text{TER}}(e_1) = 0.5, \quad P_{\text{TER}}(e_2) = 0.7, \quad P_{\text{TER}}(e_3) = 0.4$$

Step 1: Computation of Element-wise Maximum for Each Emotion

The element-wise maximum for each emotion is computed as follows:

$$P_{\max}(e_1) = \max(0.6, 0.5) = 0.6$$
$$P_{\max}(e_2) = \max(0.3, 0.7) = 0.7$$
$$P_{\max}(e_3) = \max(0.2, 0.4) = 0.4$$

Step 2: Selection of the Emotion with the Highest Probability
The emotion with the highest maximum probability is selected by:

$$e_{\text{chosen}} = \arg\max_i(P_{\max}(e_1), P_{\max}(e_2), P_{\max}(e_3))$$

In this case, the maximum values are:

$$P_{\max}(e_1) = 0.6, \quad P_{\max}(e_2) = 0.7, \quad P_{\max}(e_3) = 0.4$$

The maximum value is 0.70, which corresponds to emotion e_2 (sad). Therefore, the chosen emotion is:

$$e_{\text{chosen}} = e_2 \quad \text{(sad)}$$

The goal of the emotion fusion mechanism is to select the most dominant emotion from the combined multimodal input. In cases where the detected emotions from both modalities align, this fusion step reinforces the system's emotional understanding. If there is conflict or discrepancy between the modalities (e.g., the face shows happiness but the text indicates sadness), the fusion mechanism resolves this conflict by weighing the most likely emotional outcome based on contextual information.

The justification of the choice of a max-probability decision-based fusion mechanism can be explained as follows:

- Independence of Modalities and Supplementary Information: There is no shared feature space between the FER and TER models that enables direct feature fusion. Moreover the goal is for the models to supplement each other i.e prove to be effective in scenarios where one modality fails or both give inconsistent results. For instance, if text is sarcastic or emotionally neutral while facial expressions are expressive, the FER's output naturally dominates the fusion result and vice versa.
- Simplicity and Robustness: The max-based fusion mechanism isn't computationally expensive and proves ideal for quick inference in real-world scenarios. Having a weighted-sum fusion would require tunable parameters and a supervised multimodal dataset for the same and since the proposed framework focuses on individual models and simplicity in interpretation, a complex fusion strategy isn't employed.

This approach ensures that the assistant's emotional understanding is informed by the strongest signal from either the user's facial expression or textual content, thus offering a more nuanced and accurate emotional response.

D. LLM RESPONSE GENERATION

The response generation component in the proposed system utilizes DialoGPT, a large-scale generative pre-trained language model, to produce responses that are both contextually relevant and emotionally aligned with the user's detected emotional state. The primary objective is to generate responses that reflect the emotional nuances detected from the user's multimodal inputs (textual and visual), thereby enhancing the chatbot's empathy and engagement in real-time interactions.

This stage distinguishes our system from conventional FER-TER models. By directly connecting the predicted emotion state to a fine-tuned DialoGPT model, we ensure not just understanding, but context-aware, emotionally aligned language generation. Unlike standard intent-based chatbots, this integration displays the potential for adaptation to various LLMs, enabling dynamic multimodal interactions.

1) FINE-TUNING FOR EMOTION-AWARE RESPONSES

DialoGPT was further fine-tuned on a custom dataset specifically structured to align conversational intents (e.g., greetings, farewells) with the emotional context identified from user input. This approach ensures that the model not only recognizes the intent behind a user's message but also adapts its tone and content to match the detected emotional state, such as joy, sadness, or neutrality.

The fine-tuning dataset categorizes intents based on conversational contexts and emotions. (Fig 10) For example, for a ''hello'' intent, responses are customized based on the detected emotion: a neutral response might be ''Hello! How can I assist you?'', while a response for a user expressing joy could be ''Hey! Feeling good?'', and for sadness, ''Hi. If you need to talk, I'm here.'' This dynamic response generation based on emotional context allows the chatbot to respond in a way that feels personalized and empathetic, fostering a more human-like interaction.

2) TRAINING AND OPTIMIZATION PROCESS

The fine-tuning process for DialoGPT is optimized to support both intent recognition and emotion-conditioned responses. The model is trained to map conversational intents to specific emotional response categories using cross-entropy loss, which encourages the generation of responses that are contextually accurate and emotionally nuanced. This approach enables the model to provide coherent, emotion-aware responses that align with both the conversational flow and the user's emotional cues.

3) MUTUAL INFORMATION MAXIMIZATION FOR DIVERSE RESPONSES

To further enhance the diversity and relevance of responses, the model employs Mutual Information Maximization (MMI) scoring. This method helps rerank response candidates to prioritize those that are informative and avoid repetitive or generic replies, which can detract from the conversational quality. By balancing relevance and emotional alignment, the system achieves a high degree of conversational adaptability.

4) EVALUATION OF EMOTION-AWARE RESPONSE GENERATION

The effectiveness of the fine-tuned DialoGPT model was evaluated using a multi-reference Reddit dataset with approximately 6,000 examples. Beam search and human-annotated

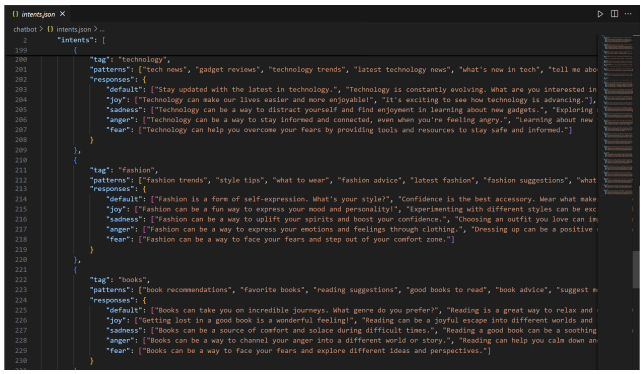


FIGURE 10. Custom Dataset for LLM fine tuning.

responses were used to benchmark the model's performance, measuring its ability to produce emotion-aware responses that are consistent with human expectations. The evaluation demonstrated that DialoGPT's responses met high standards of contextual and emotional alignment, validating its utility for real-world applications.

This structured training and evaluation process allows DialoGPT to enhance the chatbot's ability to generate emotionally intelligent responses, making it a robust tool for empathetic and context-sensitive interactions.

E. EXPERIMENTAL SETUP

To evaluate the proposed multimodal emotion detection framework, including individual models and the integrated system, the following experimental setup was used.

1) HARDWARE CONFIGURATION

- GPU : NVIDIA RTX 3060, 6GB VRAM
- CPU : AMD Ryzen 7 6800HS
- RAM : 16 GB
- Storage : 1TB SSD

2) SOFTWARE ENVIRONMENT

- Programming Language : Python 3.11.5
- Deep Learning Libraries : PyTorch 2.2.2, TensorFlow 2.16.1
- Frameworks : React, Flask

The experimental setup ensured optimal performance for both individual models and the integrated system. FER and TER models were trained on the GPU, with datasets preprocessed and loaded using TensorFlow and PyTorch pipelines. For integration, Flask APIs connected the models to the backend, where outputs were fused and passed to the fine-tuned DialoGPT for emotion-aware response generation. A React-based frontend provided a dynamic interface for real-time interaction and result visualization, ensuring seamless system usability.

IV. RESULTS AND DISCUSSIONS

The following section evaluates the multimodal emotion recognition framework that integrates FER, TER, and LLM Response Generation. On GoEmotions, the MiniLM transformer achieved 59% validation accuracy, despite issues with overlapping emotion classifications such as "neutral." Good real-time performance and 70% with a deviation of $\pm 1.03\%$ accuracy were attained by the ViT-Tiny model on AffectNet. The improved DialoGPT increased user engagement by producing emotionally connected and contextually relevant responses. Although there were minor incidents that pointed to areas that needed work, the framework has a lot of room for development and real-world applications.

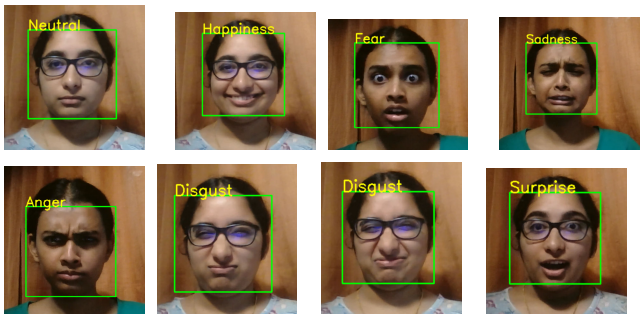


FIGURE 11. Real time classification of emotions.

A. FACIAL EMOTION RECOGNITION

The proposed FER model achieves 71% test accuracy which is competitive to 52-55% by Pourmirzaei et. al [44] and 57-58% by Li et.al [45]. The base paper for AffectNet [27] has test accuracies in the range 53-70% for individual classes and ours performs on a similar scale achieving accuracies upto 90% on certain classes. ViT-Tiny for FER is a lightweight version that retains the benefits of the full ViT model, such as capturing subtle facial features important for FER. This makes it ideal for deployment on real-time resource-constrained devices, achieving strong performance without compromising efficiency.

TABLE 5. Overall performance metrics for the FER model.

	Precision	Recall	F1-Score
macro avg	0.60	0.59	0.58
weighted avg	0.70	0.71	0.69

The model's real-time accuracy was demonstrated by combining it with OpenCV, incorporating Haar Cascades for face detection. In live scenarios, the model successfully classified facial expressions across the following labels: "Fear," "Surprise," "Anger," "Happiness," "Disgust," "Sadness," and "Neutral". This has been shown in Fig 11.

Table 6 provides the classification report, detailing precision, recall, and F1-scores for each emotion class. Fig 12 helps to visualize the same. These metrics illustrate the model's ability to capture a broad range of emotional expressions.

Its ability to perform well under real-world conditions supports its integration into the system, ensuring better

TABLE 6. Performance metrics for each emotion of the FER model.

Label	Precision	Recall	F1-Score
Anger	0.71	0.50	0.58
Disgust	0.60	0.48	0.53
Fear	0.64	0.43	0.52
Happiness	0.92	0.89	0.91
Neutral	0.73	0.94	0.82
Sad	0.67	0.98	0.80
Surprise	0.52	0.51	0.51

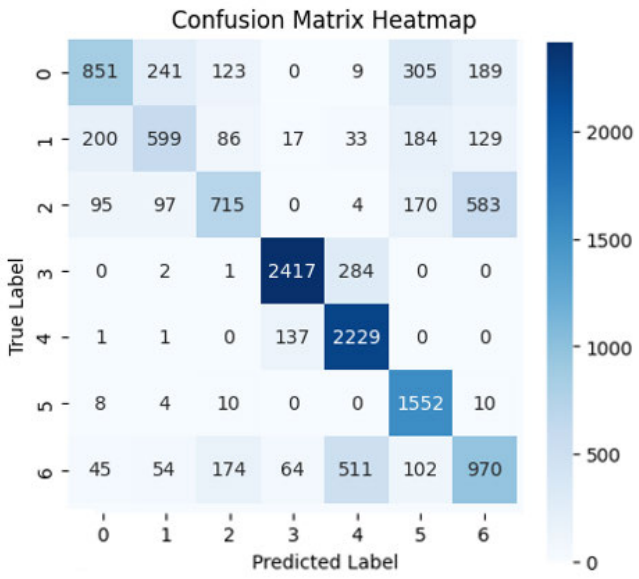
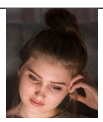


FIGURE 12. Confusion matrix for FER model.

TABLE 7. Performance of ViT tiny on low illuminated and angled images.

Image	True Label	Predicted Label
	Anger	Anger
	Happy	Happy
	Sad	Sad

generalization and more accurate emotion detection in practical applications as seen in Table 7.

B. TEXTUAL EMOTION RECOGNITION

The finalized model is the MiniLM-L12-H384-uncased model developed by Microsoft. It is a compact and efficient transformer with 12 layers, 384 hidden units, and 12 attention heads, featuring 33 million parameters, and delivers

TABLE 8. Overall performance metrics for the TER model.

	Precision	Recall	F1-Score
macro avg	0.50	0.39	0.42
weighted avg	0.58	0.59	0.58

TABLE 9. Performance metrics for each emotion of the TER model.

	Precision	Recall	F1-Score	Support
joy	0.69	0.76	0.72	3788
sadness	0.44	0.23	0.30	722
anger	0.43	0.42	0.43	1826
surprise	0.46	0.18	0.26	731
neutral	0.60	0.69	0.65	5898
disgust	0.39	0.19	0.26	306
fear	0.46	0.27	0.34	729

TABLE 10. Emotion detection for some sample real-time sentences.

Sentence	Emotion Detected
Oh my God! You came to meet me finally!	Surprise
I missed my bus	Sadness
That looks gross	Disgust
Ugh! Stop wasting my time	Anger
I don't want to be alone in the dark. It is scary	Fear
My colleague is so annoying	Anger
Stop acting as if you are the only one here	Anger
I failed my math test	Sadness

performance that is 2.7 times faster than BERT-Base while maintaining competitive accuracy.

The MiniLM model achieves a validation accuracy of 59% and delivers satisfactory real-time predictions. The GoEmotions base paper [34] reports an average F1 score of 64% for the Ekman-grouped model (7-emotion taxonomy) using BERT-base. While MiniLM falls slightly short of this, the difference is reasonable considering its significantly lighter architecture. Furthermore, the original study did not employ the efficient preprocessing pipeline used in this work, which introduces slight inconsistencies when comparing results (BERT-Base with the proposed preprocessing gives an F1-score of 56% as indicated in Table 4) Notably, there has been limited experimentation on this dataset overall. Among the other BERT-based variants evaluated (Table 3, 4), MiniLM outperforms them in accuracy. For instance, DistilRoBERTa, another lightweight alternative, achieves only 57%, reinforcing MiniLM's advantage as a compact yet high-performing model for this task.

The summary of the performance metrics of the chosen model is given in Table 8 and the class-wise performance in Table 9 which can be visualized through Fig 13. This showcases the limitations of textual emotion recognition, as the emotion classes often overlap with "neutral" in many instances. As suggested by literature, this is frequently due to the use of idioms and phrases, which cannot be interpreted at face value.

The predictions of the model with a few sample real-time sentences (not present in the dataset) is illustrated in Table 10. The model effectively detects emotions even without explicit

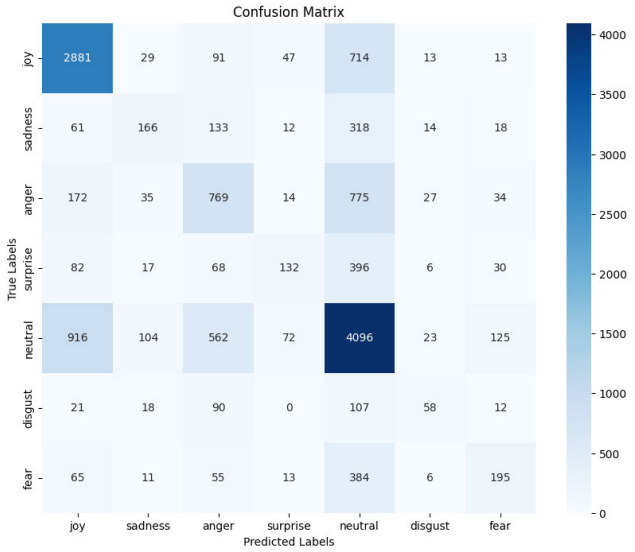


FIGURE 13. Confusion matrix for TER model.

TABLE 11. Emotion detection with sarcasm context.

Sentence	Original Emotion Detected	Context Aware Emotion Detected
Oh great! I missed the bus. Now I have to wait for hours before the next one.	Joy	Sarcasm + Anger
Pineapples on pizza. Gross.	Neutral	Sarcasm + Disgust
You're doing a great job at threatening me.	Joy	Sarcasm + Anger

cues. For example, *"I failed my math test"* implies *Sadness* through context rather than keywords. Similarly, *"Stop acting as if you are the only one here"* reflects *Anger* without directly stating it. This suggests the model leverages contextual patterns, tone, and phrasing learned from data, allowing it to infer emotions beyond surface-level sentiment words.

Table 11 highlights the model's performance with a few sentences that have obvious sarcastic cues. Phrases like *"Oh great! I missed the bus"* are likely to be misclassified as *Joy* due to positive phrasing. However, with context-aware analysis, the model correctly identifies *Sarcasm + Anger*, showing its ability to detect emotional undertones masked by irony. This demonstrates the effectiveness of integrating the sarcasm-detection module in the enhancement of the pipeline.

C. LLM RESPONSE GENERATION AND USER INTERFACE

The fine-tuned DialoGPT model performed well in generating contextually relevant and emotionally aligned responses, validated by both automated metrics and human evaluations. A sample response of the model is depicted in Fig 14.

The model effectively adjusted its tone according to the detected emotion. For instance, responses were empathetic when sadness was detected and upbeat when joy was present, enhancing user engagement. An example is portrayed in Fig 15 where the user asks for a song suggestion and does not specify any additional information. The virtual assistant detects the user's emotion from the video feed as 'happy' and suggests an appropriate song thus enhancing the user experience.

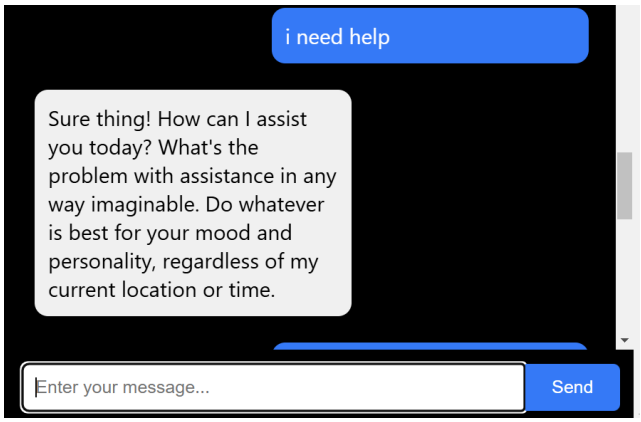


FIGURE 14. Sample screenshot of LLM response generation when asked for help.

Various emotion combinations and responses are tested out as demonstrated in Fig 16. The chatbot replies according to the most prominent emotion based on the probability. For example, in the first image, it is clearly evident that the user is disgusted from their text even though their face remains relatively impassive with a slight percentage of surprise detected. The other images depict similar situations, highlighting the creative response generation feature of the model. Similarly, Table 12 provides a more detailed understanding of how the two modalities can supplement each other to create a comprehensive interpretation of the user's emotional state. This enables the system to identify the most prominent emotion and generate more accurate, contextually appropriate responses.

Overall, DialoGPT generated varied responses, avoiding the blandness common in many chatbots. In ambiguous cases, emotional alignment was occasionally less precise. Further refinement with diverse, emotion labeled data could improve nuanced response generation.

D. CONTRASTS WITH TRADITIONAL MER SYSTEMS

MER systems that rely on large-scale datasets face several critical challenges, including dataset annotation complexity, multimodal fusion difficulties, and generalization issues.

Traditional MER models, such as those based on CMU-MOSEI, IEMOCAP, and MELD, often struggle with the inherent subjectivity in emotion labeling, which introduces inconsistencies in training. These datasets are collected in controlled environments, where interactions are often scripted or semi-scripted. This reliance on non-realistic scenarios limits the adaptability of these models to real-world

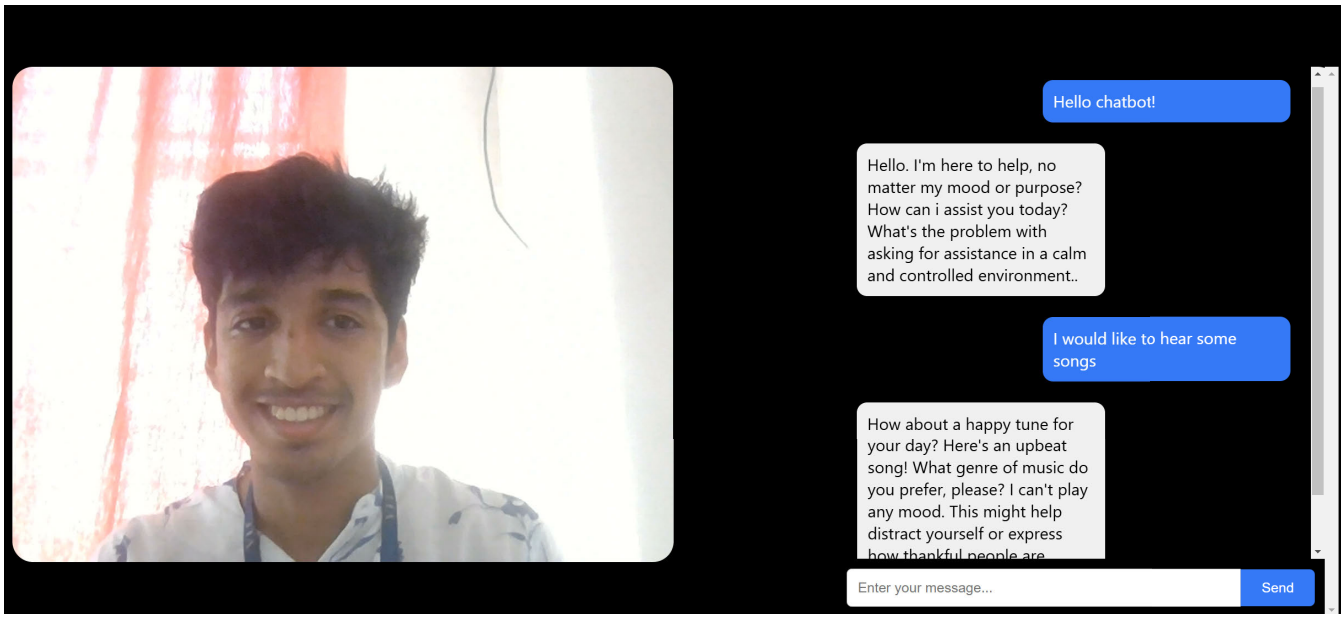


FIGURE 15. Virtual Assistant suggesting a happy song based on the emotion deduced from the video feed.

TABLE 12. Demonstration of FER and TER as supplementary modalities to improve emotion understanding in a multimodal system.

Prompt	FER Only Response	TER Only Response	FER + TER Response	Comments
Oh no! I missed my bus....Can you suggest some music?	(Neutral Face) Detected Emotion: Neutral Music is a great way to unwind and reset, away from the sources of anger or frustration. How about 'Break Stuff' by Limp Bizkit?	Detected Emotion: Sadness Here's a cheerful song to lift your spirits! How about 'Can't stop the Feeling!' by Justin Timberlake?	Selected Emotion: Sadness How about 'Fix You' by Coldplay? It's got energy and it might help improve your mood!	TER captures the emotion better. FER misinterprets due to a neutral face. The combined system uses TER's emotion for the response.
Suggest me a song.	(Happy Face) Detected Emotion: Happy Upbeat music can amplify the feeling of happiness. How about 'Weightless' by Justin Timberlake? Its energy might resonate with your current feelings!	Detected Emotion: Neutral Music can help you relax and unwind after an intense day? How about 'Sunday Morning' by Maroon 5.	Selected Emotion: Happy Oh, any positive music can help improve your mood. How about 'Weightless' by Macaroni Union? I'm feeling good! Let's play some upbeat, even cheerful songs or artists you love!	FER correctly identifies happiness, while TER lacks emotional context. The mulitmodal system takes up the FER detected emotion for emotional context.

applications, as they struggle with spontaneous and diverse emotional expressions encountered in industrial or open-world settings.

In contrast, FER and TER datasets, such as AffectNet and GoEmotions, are derived from more naturally occurring data. AffectNet, for instance, contains facial expressions sourced from the internet, representing a broader range of real-world emotions, while GoEmotions is built on user-generated content from social media, capturing spontaneous textual expressions of emotion. These datasets provide a more diverse and representative emotional distribution compared to traditional multimodal datasets, which are often constrained by actor-performed expressions or scripted dialogues.

As a result, MER models, despite their sophisticated architectures, often suffer from poor generalization when applied to unseen domains or real-world interactions. The heterogeneity in data sources across modalities further complicates the fusion process, making it difficult to align multimodal signals effectively. Our proposed framework mitigates some of these challenges by leveraging FER and TER models trained on diverse datasets, ensuring better generalization across different emotional contexts. By integrating a ViT-Tiny-based FER model with 70% accuracy on AffectNet, a MiniLM-based TER model with 59% accuracy on GoEmotions, and DialoGPT for generating emotionally resonant responses, our approach balances efficiency with adaptability.

Ultimately, while existing MER models excel within their respective dataset constraints, their transferability to real-world applications remains a challenge. Addressing

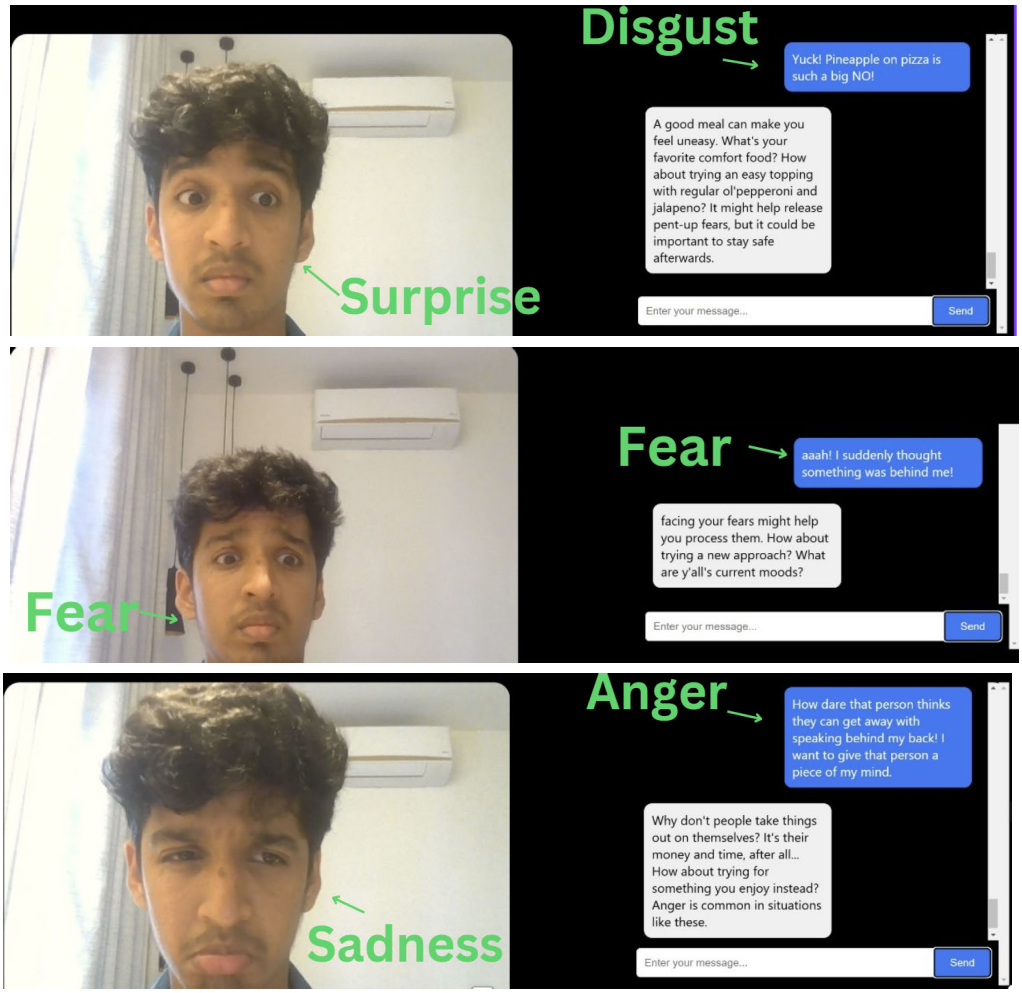


FIGURE 16. Varying Responses for Different Categories of Emotions Detected.

the construction of more universal and domain-adaptive MER models is crucial for expanding their usability beyond controlled research environments. Our framework takes a step in this direction by utilizing datasets that better reflect spontaneous emotional expressions, improving real-world applicability while maintaining computational efficiency [46].

V. CONCLUSION

This study introduced a multimodal approach for emotion recognition that combines FER, TER, and response generation using a fine-tuned DialoGPT model. By integrating FER based on a ViT architecture with TER using the MiniLM model, the system can classify user emotions in real-time and generate emotionally aligned responses. Our FER model, trained on the AffectNet dataset, demonstrated robust real-time performance across diverse emotions, while the TER model achieved a validation accuracy of 59%. The combined output of these models allowed the DialoGPT model to produce responses that matched the emotional tone of the user, adding empathy and engagement to interactions. The intents dataset on which DialoGPT was trained upon, provides us with a gateway for customizations of conversation based on the application of this system. For this study, the intents were kept generalized to facilitate simple conversations. Application specific intents could be curated to ensure conversations are audience appropriate and relevant to the topic at hand. This also ensures better engagement with the users.

Overall, this work contributes a scalable and effective framework for building emotionally intelligent virtual assistants. With further improvements in multimodal fusion and response personalization, the system has strong potential for applications in customer support, therapy, and other emotion-sensitive domains, advancing the goal of creating conversational agents that interact with users in more human-like, empathetic, and engaging ways.

VI. FUTURE WORKS

Improving the individual model performances is a possible aspect of future work. Currently the performance of the

models are at par with the benchmarks for the individual datasets.

For TER, multi-task learning and data augmentation techniques like paraphrasing or producing synthetic data can be experimented with. Additionally, handling complex emotions like sarcasm and mixed emotions as well as improving contextual understanding by incorporating conversation history will increase accuracy. Using ensemble techniques, such as transformer-based models in conjunction with CNNs or LSTMs and cross-model transfer learning, can further enhance the model's performance.

For FER, occlusions—such as hands, hair, or objects partially covering the face—can obscure important facial features, often leading to incorrect or uncertain emotion predictions. While the model still attempts to classify emotions based on the visible regions, these situations typically result in less reliable outcomes. To reduce the risk of misinterpreting emotions in such cases, a potential solution is to incorporate a face and landmark detection system, like Dlib's HOG-based detector. This approach identifies and maps facial landmarks, making it possible to detect when parts of the face are occluded or missing. Once flagged, these instances could either be marked as low-confidence predictions deferring to another modality such as text. This layer of detection can enhance the reliability and transparency of FER systems, particularly in real-world scenarios where clean, unobstructed facial images cannot always be guaranteed. The addition of intensity labels can be explored which may allow the model to recognize emotions with varying expressiveness levels. Variations in head pose and occlusions can be addressed by 3-D face models or pose normalization. Additionally, the integration of self-supervised contrastive learning techniques to enhance the robustness of facial emotion representations, especially in semi-supervised or low-resource settings as described in [47] can be explored.

Emotion detection can be further enhanced by including audio as an additional input. To help understand emotions, Speech Emotion Recognition (SER) models can extract prosodic features like pitch and tone. A multimodal fusion technique that integrates text, audio, and facial recognition data will result in a more comprehensive emotion recognition system. This approach can improve empathy and context awareness for sentiment-aware audio chat-bots and other real-time, multimodal systems that process text, audio, and facial inputs simultaneously.

With respect to the emotion fusion mechanism, adaptive attention-based fusion strategies can be explored to enhance the predictions and quantify them. A unified multimodal dataset can be employed and the contribution of the individual modalities can be quantitatively evaluated. Implementing uncertainty estimation techniques, such as Monte Carlo Dropout or Gaussian Process Regression, can further enhance the robustness of emotion recognition [48]. Furthermore, self-supervised learning techniques could be leveraged to refine cross-modal embeddings, ensuring optimal information retention even under suboptimal conditions.

Future directions for the Large Language Model (LLM) include improving its ability to identify and react to emotions detected from other modalities, such as text, audio, and facial expressions. In order to create more sympathetic interactions, this might entail providing the LLM with a larger dataset of conversations that have been labeled with emotions. Enhancing human-AI interactions in customer service, therapy bots, and other applications that demand emotional intelligence would involve fine-tuning the model to modify its tone, formality, and response style based on the identified emotion. Furthermore, adding feedback systems for ongoing interaction-based learning may eventually make the LLM more customized and adaptive. The development of highly interactive and emotionally intelligent conversational agents will be made possible by further research into multimodal LLMs that can process and react using text, audio, and visual data.

STATEMENTS AND DECLARATIONS

FUNDING

No funding was received for conducting this research.

CONFLICTS OF INTEREST

The authors do not have any competing interests related to the content of this article to disclose.

AVAILABILITY OF MATERIAL AND DATA

The datasets used in this study are publicly available. Specific details and access links are included in the references.

CODE AVAILABILITY

The code related to this work is not publicly available during the review phase. If the reviewer requires access to the code, we will provide it without delay.

REFERENCES

[1] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 49–67, Jan. 2023.

[2] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput. Appl.*, vol. 35, no. 32, pp. 23311–23328, Nov. 2023.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[4] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 5776–5788.

[5] F. Schiavo, L. Campitiello, M. D. Todino, and P. A. Di Tore, "Educational robots, emotion recognition and ASD: New horizon in special education," *Educ. Sci.*, vol. 14, no. 3, p. 258, Mar. 2024.

[6] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DialoGPT: Large-scale generative pre-training for conversational response generation," 2019, *arXiv:1911.00536*.

[7] A. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions," *Inf. Fusion*, vol. 105, May 2024, Art. no. 102218.
[8] N. Ahmed, Z. A. Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intell. Syst. Appl.*, vol. 17, Feb. 2023, Art. no. 200171.
[9] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
[10] B. Pan, K. Hirota, Z. Jia, and Y. Dai, "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, vol. 561, Dec. 2023, Art. no. 126866.
[11] K. Machová, M. Szabóova, J. Paralič, and J. Mič ko, "Detection of emotion by text analysis using machine learning," *Frontiers Psychol.*, vol. 14, Sep. 2023, Art. no. 1190326.
[12] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
[13] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion detection in text: A review," 2018, *arXiv:1806.00674*.
[14] H. T. Phan, D. T. Pham, and N. T. Nguyen, "FeDN2: Fuzzy-enhanced deep neural networks for improvement of sentence-level sentiment analysis," *Cybern. Syst.*, vol. 2023, pp. 1–17, Dec. 2023.
[15] B. Maleszka, "How to explain sentiment polarity – a survey of explainable sentiment analysis approaches," *Cybern. Syst.*, vol. 2023, pp. 1–17, Dec. 2023.
[16] R. Harika, T. Uday, M. L. Sirisha, M. S. L. Sahitya, K. Drugaanjali, and M. S. Srinivas, "A review of advancements in facial emotion recognition and detection using deep learning," in *Proc. Int. Conf. Social Sustain. Innov. Technol. Eng. (SASI-ITE)*, Feb. 2024, pp. 290–295.
[17] U. Mishra, "Emotion detection in virtual assistants and chatbots," *Int. Res. J. Modernization Eng. Technol. Sci.*, 2023, doi: 10.56726/IRJMETS44626.
[18] G. Bilquise, S. Ibrahim, and K. Shaalan, "Emotionally intelligent chatbots: A systematic literature review," *Human Behav. Emerg. Technol.*, vol. 2022, pp. 1–23, Sep. 2022.
[19] M. Abuhmida, M. J. Islam, and W. Booth, "Empathy in ai: Developing a sentiment-sensitive chatbot through advanced natural language processing," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 13, no. 3, pp. 1–10, 2024.
[20] W. Wang, X. Cai, C. Hsuan Huang, H. Wang, H. Lu, X. Liu, and W. Peng, "Emily: Developing an emotion-affective open-domain chatbot with knowledge graph-based persona," 2021, *arXiv:2109.08875*.
[21] D. Griol, A. Sanchis, J. M. Molina, and Z. Callejas, "Developing enhanced conversational agents for social virtual worlds," *Neurocomputing*, vol. 354, pp. 27–40, Aug. 2019.
[22] P.-S. Chiu, J.-W. Chang, M.-C. Lee, C.-H. Chen, and D.-S. Lee, "Enabling intelligent environment by the design of emotionally aware virtual assistant: A case of smart campus," *IEEE Access*, vol. 8, pp. 62032–62041, 2020.
[23] K. Bahreini, R. Nadolski, and W. Westera, "Towards multimodal emotion recognition in e-learning environments," *Interact. Learn. Environ.*, vol. 24, no. 3, pp. 590–605, Apr. 2016.
[24] R. Duwenbeck and E. A. Kirchner, "Auditive emotion recognition for empathic AI-assistants," *KI - Künstliche Intelligenz*, vol. 38, no. 3, pp. 151–156, Nov. 2024.
[25] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.
[26] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, vol. 64, Apr. 2015, pp. 59–63.
[27] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.
[28] A. Nouisser, R. Zouari, and M. Kherallah, "Enhanced MobileNet and transfer learning for facial emotion recognition," in *Proc. Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2022, pp. 1–5.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[30] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2019, pp. 6105–6114.
[31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2001.
[33] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Jan. 2018, pp. 3687–3697.
[34] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," 2020, *arXiv:2005.00547*.
[35] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0144296.
[36] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 101–111, Apr. 2014.
[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[38] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proc. 1st Instruct. Conf. Mach. Learn.*, 2003, pp. 29–48.
[39] S. M. Mohammad, "Practical and ethical considerations in the effective use of emotion and sentiment lexicons," 2020, *arXiv:2011.03492*.
[40] S. Buechel and U. Hahn, "Emotion representation mapping for automatic lexicon construction (Mostly) performs on human level," 2018, *arXiv:1806.08890*.
[41] A. I. Kadhim, "An evaluation of preprocessing techniques for text classification," *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018.
[42] C. P. Chai, "Comparison of text preprocessing methods," *Natural Lang. Eng.*, vol. 29, no. 3, pp. 509–553, May 2023.
[43] M. Romero. *T5-base Fine-tuned for Sarcasm Detection on Twitter*. Accessed: May 2025. [Online]. Available: https://huggingface.co/mrm8488/t5-base-finetuned-sarcasm-twitter
[44] M. Pourmirzaei, G. Ali Montazer, and F. Esmaili, "Using self-supervised auxiliary tasks to improve fine-grained facial representation," 2021, *arXiv:2105.06421*.
[45] J. Li, J. Nie, D. Guo, R. Hong, and M. Wang, "Emotion separation and recognition from a facial expression by generating the poker face with vision transformers," *IEEE Trans. Computat. Social Syst.*, early access, Nov. 5, 2024, doi: 10.1109/TCSS.2024.3478839.
[46] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, p. 1440, Oct. 2023.
[47] B. Fang, X. Li, G. Han, and J. He, "Rethinking pseudo-labeling for semi-supervised facial expression recognition with contrastive self-supervised learning," *IEEE Access*, vol. 11, pp. 45547–45558, 2023.
[48] R. Zhao, K. Wang, Y. Xiao, F. Gao, and Z. Gao, "Leveraging Monte Carlo dropout for uncertainty quantification in real-time object detection of autonomous vehicles," *IEEE Access*, vol. 12, pp. 33384–33399, 2024.

SHAUN GEORGE RAJESH is currently pursuing the Bachelor of Technology degree in computer science and engineering with Vellore Institute of Technology, Chennai. With a deep passion for artificial intelligence, deep learning, computer vision, and natural language processing, he has been actively exploring these fields since his first year on campus. He has worked on several innovative projects and gained practical experience in cutting-edge technologies. He was a Finalist in the Prestigious National-Level Dark Pattern Buster Hackathon 2024, showcasing his problem-solving skills and creativity. In addition to his technical expertise, he is also passionate about emceeing and has demonstrated strong team leadership abilities, excelling in collaborative environments.

SMRITI VIPIN MADANGARLI is currently pursuing the bachelor's degree in computer science and engineering with Vellore Institute of Technology, Chennai. She is particularly interested in core programming, algorithm design and implementation, and software engineering. She is proficient in Python, which she has applied extensively in her research work. She also gained hands-on experience during internships, where she honed her problem-solving skills and contributed to various software development projects. Her research interests include computer vision, natural language processing, and computer systems and architectures.

GAURI SANTOSH PISHARADY is currently pursuing the bachelor's degree in computer science and engineering with Vellore Institute of Technology, Chennai. She has a strong command of Python, which she leverages extensively in projects involving machine learning and logical problem-solving. Her work demonstrates a keen ability to design and implement algorithms that address real-world challenges, particularly in areas like computer vision and game development. Through her internship experience, she has gained valuable industry exposure, applying her technical skills to develop efficient, scalable solutions and working on collaborative teams to tackle complex problem statements. Her interest in ML enables her to build predictive models and intelligent systems, combining logical reasoning with technical proficiency to deliver impactful outcomes.

ROLLA SUBRAHMANYAM received the M.Tech. degree in artificial intelligence and the Ph.D. degree in computer science from the University of Hyderabad, India. He is currently a Senior Assistant Professor with Vellore Institute of Technology, Chennai. His research interests include cryptography, artificial intelligence (AI), and the intersection of AI and cryptography.