Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multimodal Emotion Recognition: A Tri-modal Approach Using Speech, Text, and Visual Cues for Enhanced Interaction Analysis

Gauraangi Praakash[1], Dr. Pooja Khanna[2]

1 Department of Computer Science and Engineering Amity School of Engineering and Technology Lucknow, Amity University Uttar Pradesh Lucknow, India gauraangi25@gmail.com

2Department of Computer Science and Engineering Amity School of Engineering and Technology Lucknow, Amity University Uttar Pradesh Lucknow, India pkhanna@lko.amity.edu

ARTICLE INFO

Received:18 Dec 2024

Revised: 16 Feb 2025

Accepted:26 Feb 2025

ABSTRACT

In an age dominated by the rapid development of human-computer interaction, knowledge of the user emotions has become a critical building block in creating engagement as well as responsiveness. This paper presents a tri-modal system towards real-time emotion perception through the merging of textual, visual, and audio information. Our method utilizes strong deep learning networks in each of the three modalities: DistilBERT to perform sentiment analysis on text data (on the SST-2 dataset), ViT (Vision Transformer, vit-base-patch16-224-in21k) to detect emotions from faces, and a task-specific Convolutional Neural Network on the RAVDESS dataset to perform emotional analysis from speech. Each modality is processed independently and fused subsequently to get a global emotion score, thereby enabling fine-grained behavioral analysis in real-time. Tri-modal fusion not only enhances accuracy but also yields robustness against varied scenarios, thereby solving the problems due to incomplete or ambiguous information in any one of the individual modalities. The paper here shows that a uniform framework performs substantially better than unimodal systems in extracting emotional context, thereby paving the way for more intelligent as well as emotionally responsive applications in mental health monitoring, customer support, and human-robot interaction.

Keywords: emotionally, unimodal, individual, modalities

1. INTRODUCTION

Emotion recognition has now emerged as an essential field of affective computing, greatly enhancing human and intelligent system interaction. With AI-driven agents such as digital assistants, chatbots, and social robots going far and wide, the need for such intelligent systems to recognize and respond to human emotions has become increasingly important. Emotions are not only expressed through words but also through facial expressions and voice tones, and the use of a multimodal approach to effective emotion detection becomes essential. While unimodal approaches based on conventional practices may work in some cases, they are not effective in cases where the emotional cues are incomplete or ambiguous. This research aims to overcome this limitation by proposing a tri-modal emotion recognition framework combining text, visual, and audio modalities for richer analysis.

This study utilizes three deep learning models with the latest techniques applied to every modality. For text sentiment analysis, the computationally lighter version of BERT, DistilBERT, is fine-tuned on the Stanford Sentiment Treebank (SST-2) corpus. For vision emotion detection, the Vision Transformer model (ViT-base-patch16-224-in21k) is used, relying on self-attention mechanisms for efficient spatial feature extraction from facial expressions. For the audio modality, a custom Convolutional Neural Network (CNN) architecture is used, trained on the RAVDESS dataset, with a few convolutional layers

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and dropout components to support strong generalization. Each modality is processed independently and the results are later aggregated to produce an overall sentiment score.

The primary objective of the present study is to show how a tri-modal framework exhibits superior performance and context sensitivity when compared to unimodal and bi-modal systems. This holistic methodology allows the model to understand a more holistic view of emotional expression, especially for real-time environments where users may express messages via a mixture of cues. The manuscript then presents relevant literature for emotion recognition, the structural formulation of each model, the methodology used for preprocessing the dataset, the data fusion strategies, and the results obtained from the conducted experiments. The proposed framework not only contributes to the field of multimodal emotion recognition but also provides the foundations for real-world applications in the healthcare, educational, virtual communication, and user experience enhancement fields.

2. LITERATURE REVIEW

Emotion recognition has garnered substantial attention in artificial intelligence, human-computer interaction, and affective computing due to its importance in enabling machines to understand and respond to human affective states [1]. Traditionally, emotion recognition systems relied on single modalities such as facial expressions, speech, or text [2]. However, the complexity and subtlety of human emotions often require the integration of multiple modalities to achieve robust recognition [3]. Multimodal emotion recognition, which combines information from different channels, has emerged as a strong alternative to unimodal approaches [4], [5]. By fusing complementary cues from speech, text, and visual data, these systems can provide a more complete representation of emotional states [6]. This aligns with how humans perceive emotion—through vocal intonation, facial expressions, body language, and language content [7]. The convergence of signals into multi- and amodal representations further mirrors human-like emotion perception [8].

Recent advancements in textual analysis have significantly improved emotion recognition capabilities. The emergence of transformer-based models such as BERT has allowed systems to capture semantic and syntactic information from text more effectively [9]. These models have demonstrated an ability to discern subtle emotional cues in both written and spoken language. Traditional lexicon-based sentiment analysis was limited in understanding context, but with pre-trained language models like BERT and its variants, emotion understanding has reached new levels of depth and accuracy [10], [11]. Meanwhile, recurrent models such as LSTM and GRU were effective for temporal emotion modeling in speech and textual sequences, though they are increasingly being replaced by transformers due to their parallelism and attention mechanisms [12].

Multimodal fusion strategies such as early fusion, late fusion, and hybrid approaches have further improved system performance by leveraging cross-modal information exchange [13]. For instance, CNNs and ResNet variants are widely used for extracting visual features, while transformers like ViT have demonstrated effectiveness in vision-based emotion recognition tasks [14]. Datasets such as IEMOCAP, MELD, and CMU-MOSEI have provided high-quality multimodal data, fueling benchmark evaluations [15]. The combination of lightweight models such as DistilBERT with MobileViT offers opportunities for real-time inference, especially on edge devices. Moreover, explainability techniques such as attention heatmaps and saliency visualizations are gaining importance to enhance trust in emotion-aware AI systems.

3. METHODOLOGY

This research employs a tri-modal deep learning pipeline for human emotion detection by analyzing the input from three modalities, i.e., textual, visual, and audio data. All three modalities are analyzed using dedicated models and techniques that are best suited to extract emotion-carrying features from each form of input. The entire pipeline is comprised of numerous key steps such as data acquisition, preprocessing, model construction, training, testing, and fusion. Such a modular structure makes sure each module contributes effectively in the end emotion classification task and hence the final accuracy and robustness for various types of inputs. Fig.1. shows a tri-modal emotion recognition framework, with the DistilBERT processing text data, the ViT processing visual data, and CNN dealing with audio

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

data, then fusing them all together to yield a global emotion score. This combination increases the accuracy of emotion detection by merging the complementary features of each modality.

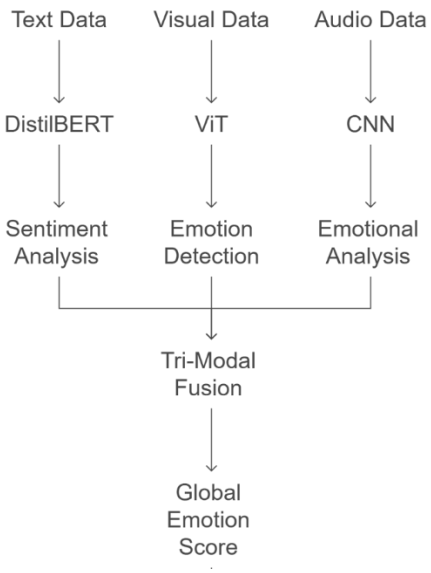

Fig.1: Architecture for Multimodal Emotion Recognition using Text, Visual, and Audio Inputs.

3.1 Data Collection and Preprocessing

The purport of this research largely lies on multimodality data collection and preprocessing. In text modality, publicly available sentiment analysis datasets with labeled textual input have been highly utilized. These datasets typically include user reviews and social media posts annotated with sentiment classes like positive, negative, or neutral. Preprocessing methods for text followed standard natural language processing (NLP) techniques such as lowercasing, and removal of punctuation, stop words, and special characters. By doing so, it ensured that the model is trained on clean, uniform text without irrelevant noise.

The human facial expression videos were used as the modalities for facial recognition. Each video was inspected in order to extract frame specific metadata such as frames per second (FPS), resolution (height and width), and total number of frames. This set of videos was then used as input for assistance in emotion recognition. Extraction was done on each frame resulting in the identification of facial regions with drawn bounding boxes that contained the localized face. This way, it ensured accurate per-frame emotion detection in the later stage.

The audio modality made use of live speech input from the user in real time. The captured audio went through preprocessing noise reduction and format standardization, critical for transcription and translation model compatibility. The three pipelines hence delivered the same noncontaminated, structured, and processed-ready data for further processing by deep learning models.

3.2 Text Modality: DistilBERT for Sentiment Analysis

A very important modality in text understanding is related to the user's sentiments through text input. The processed cleaned dataset first underwent tokenization using a session-specific pre-trained tokenizer to the DistilBERT model. Tokenization involves transforming the input sentences into sub-word units so that the model can process words outside the vocabulary.

Tokenized inputs were then preprocessed to the required input structure for DistilBERT model training. Before performing the training, the model was checked if at least one trained model can be found. If not, it would load and initialize the DistilBERT model. DistilBERT is lightweight and a faster version of BERT, and it has adopted itself to the task of competitive delivery to those of performance in classification. It was finally built, trained over the tokenized dataset, along with the sentiment labels.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

During this process-learning phase, several hyper-parameters, namely learning rate, batch size, and number of epochs, were experimented to determine their optimal performances. Fig.2. depicts the process for analyzing sentiments in text with DistilBERT. It points to key steps, ranging from data loading and preparation to model training and testing. If the trained model performs suitably, it is saved and used to generate predictions on tokenized inputs.

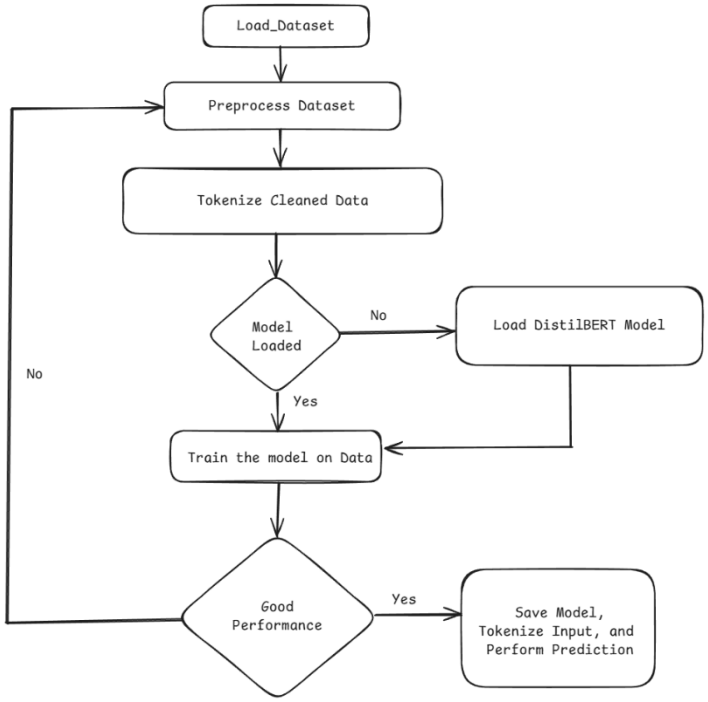

Fig.2. Flowchart for Sentiment Analysis

After training, the model's performance was evaluated using such metrics as accuracy, precision, recall and F1 scores. If the performance threshold is met, the trained model shall be saved for future usage. The model is set to be used directly for real-time tokenization of fresh input text, followed by the generation of output sentiment labels, either 'positive', 'negative', or 'neutral'. All this made for a very efficient and strong pipeline for sentiment classification based on text.

3.3 Visual Modality: Emotion Recognition from Video

In this section of visual modality, emotion-related parameters are extracted from video data. The process begins with user input which includes the path of the video file and various flags to allow visualization and output saving. The model for visual emotion recognition works with Vision Transformer (ViT) architecture called vit-base-patch16-224-in21k favorably working for image and frame-level classification tasks.

Fig.3. shows the workflow of the suggested video-based emotion recognition system. It begins with user input, loads a ViT-based model, processes video frames, labels each as predicted emotions, and outputs the dominant emotion at the end after processing all frames.Once the video path is inputted, the video is first processed to extract key parameters like frame rate (FPS), width, height, and total frame counts. These parameters are very important in ensuring correct analysis and visualization. Afterward, the video is processed in a frame-wise manner. For each of the frames, a rectangle is drawn either around detected faces or around the areas of interest, with emotion labels predicted and overlayed.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

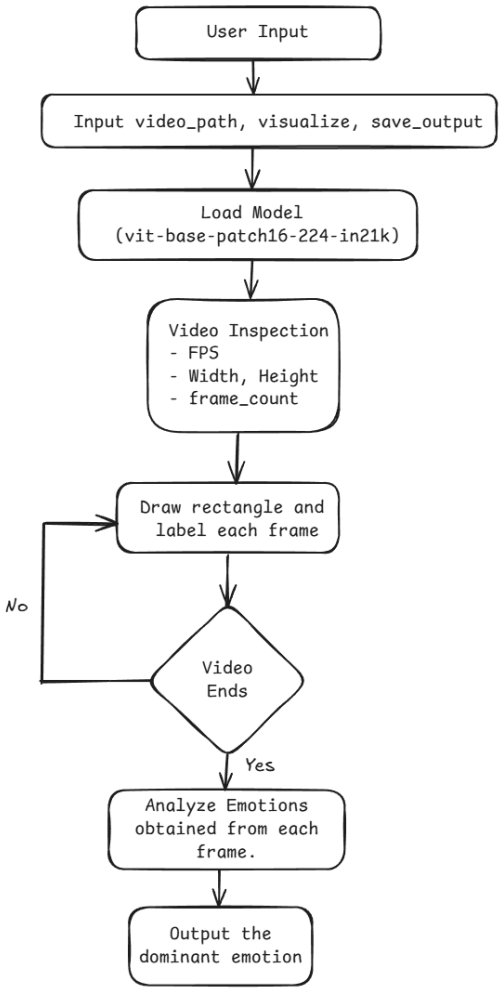

Fig.3: Flowchart of Video-Based Emotion Detection Process

The system keeps processing until the end of the video; this is when emotions are aggregated based on each frame's prediction. A dominant emotion is determined depending on either the number of predictions or the confidence of predictions within individual frames; this emotion is then considered the dominant visual emotion detected in the video.

This method ensures that the dynamic nature of emotional expressions through time is fully captured, thus rendering the visual modality a formidable partner to the other channels like text and speech in multimodal emotion recognition.

3.4 Audio Modality: Speech Emotion Analysis

Deep learning for emotional recognition works by analyzing the audio modality that extracts and interprets emotional expressions from speech. It all begins with the extraction of the audio from the original video using various tools such as moviepy or ffmpeg. After extraction, the audio file will be formatted correctly (i.e., it should be in .wav format) according to the requirements of feature extraction.

The audio is then processed to extract acoustic features such as mel-frequency cepstral coefficients (MFCCs), chroma features, zero-crossing rate, or spectral contrast, using libraries such as librosa; these capture a multitude of features reflecting changes in pitch, tone, energy, and rhythm relating to various emotional states and hence finds a lot of application in emotion recognition.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

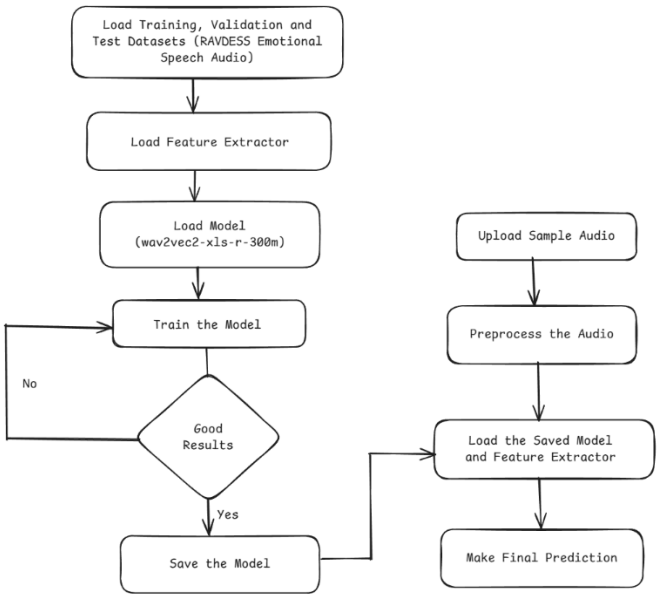

Fig.4. Workflow of Audio Emotion Recognition

Fig.4. shows the flowchart for working of speech emotion recognition.The extracted features are fed into any deep-learning algorithms that have been trained on a labeled speech emotion dataset. The architectures vary from CNNs to RNNs or even hybrid architectures aimed at modeling the temporal patterns of speech. The learned model predicts the probabilities for each emotion class, and the class with the highest predicted probability is regarded as detected emotion. This emotion will, in conjunction with predictions from other modalities, enter the storage for late fusion into the overall emotion assessment. Through the analysis of prosodic and acoustic cues, the audio modality greatly aids in sensing nuanced and very expressive emotional patterns that sometimes are not even visible and cannot be expressed linguistically.

3.5 Multimodal Fusion Strategy

The strength of the proposed system lies in the integration of predictions from all three modalities, text, visual, and audio, using an efficient multimodal fusion strategy. In this case, the modalities provide complementary information: text reveals linguistic sentiment; visuals inform facial expressions; and audio refers to the tone of voice. Late fusion provides a useful means to combine these complementary cues.

Accordingly, in late fusion, the predictions from each modality were obtained independently in the form of probability distributions or categorical labels and these predictions were then combined together using majority voting, weighted averaging, or some simple rule-based logic. In this research, we adopted a weighted-average approach whereby each modality is given some weight on the basis of its confidence derived from its historical performance and reliability across different contexts. For instance, when the facial cues are very clear, they can be given a higher weight, whereas, in the case of subtle vocal tones, text would push the defined weights in their favor, especially when sentiments expressed in the speech are very strong.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

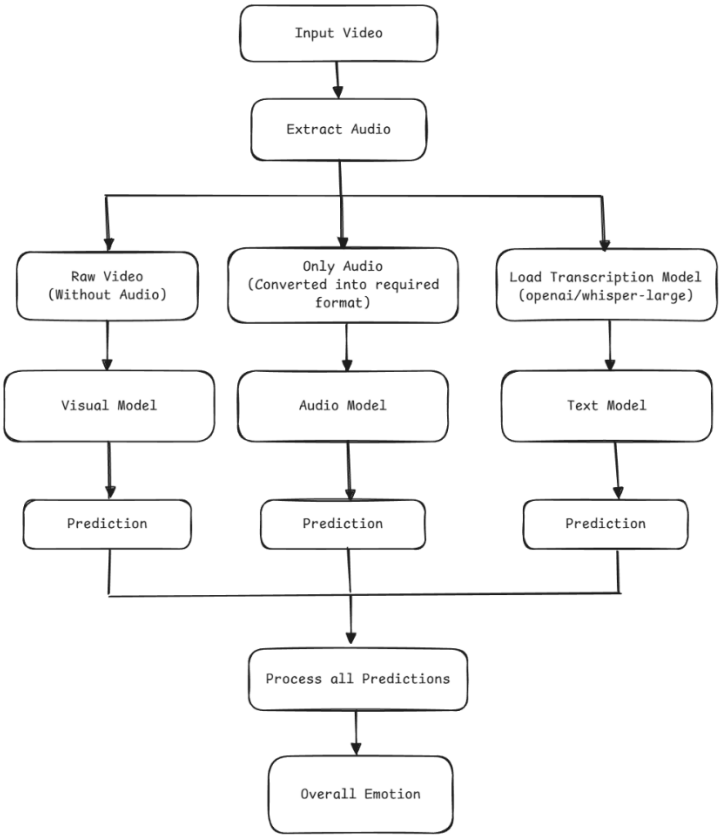

Fig.5. Multimodal Emotion Recognition Pipeline Integrating Visual, Audio, and Textual Cues from Video Input

Fig.5. illustrates a late fusion-based multimodal emotion recognition system that processes video to extract and analyze visual, audio, and text data independently. The final emotion prediction results from the weighted-average fusion of the outputs from the three modalities for enhanced robust and accurate classification. Thus the final emotion label can be assigned after analyzing the overall scores or votes of all modalities. This method utilizes the strength of each modality and compensates for the individual weaknesses, resulting in a more accurate and robust emotion recognition framework. Besides improving accuracy, such fusion also serves to improve resistance to noise or ambiguity in any single input source.

4. RESULTS AND DISCUSSION

The multimodal emotion recognition system under discussion was assessed concerning individual modality accuracy and the effectiveness of the fusion strategy. Each convective model-text, visual, or audio-was trained and tested separately before output integration.

Class	Precision	Recall	F1-Score	Support
negative	0.89	0.75	0.82	2431
positive	0.59	0.80	0.68	1103
accuracy			0.77	3534
macro avg	0.74	0.77	0.75	3534
weighted avg	0.80	0.77	0.77	3534

Table 1: Performance Metrics of DistilBERT-Based Sentiment Classifier on Text Modality

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The DistilBERT-trained sentiment classifier adapted to the text modality attained commendable performance on sentiment classification tasks. Table 1 indicates that this model achieved relatively high precision (0.89) in detecting the negative class and quite good recall (0.80) in predicting the positive class, therefore demonstrating the system's balanced efficiency across both sentiment polarities. The overall performance measured by accuracy of 77%, while agglomerate macro and weight average point estimators reflect more or less equal nature among classes. These results therefore suggest good generalization, especially for pointing out quite subtle contextual clues in the textual domain, although there's a bit of leeway to improve precision in detecting positive sentiment through further fine-tuning.

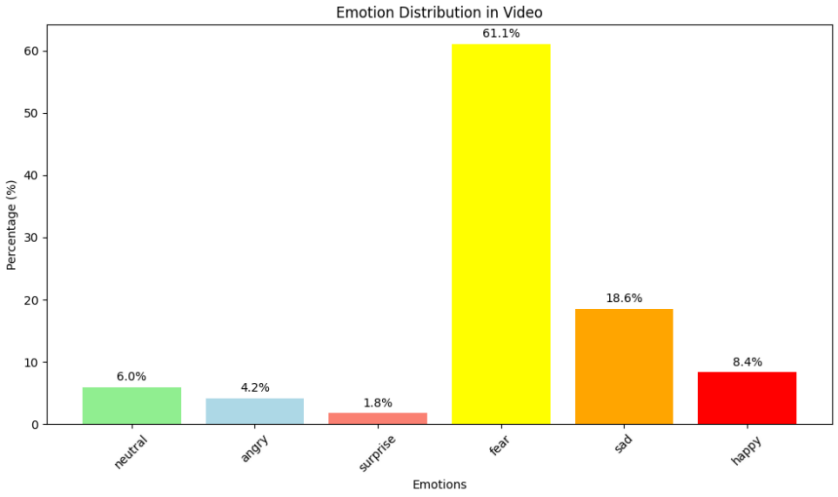

Fig.6. Emotion Distribution Detected in Video Using Emotion Recognition System

The bar chart shows the frequency of different emotions detected in the analyzed video. 'Fear' is the most dominant emotion, comprising 61.1% of the total emotional content, followed by 'Sad' at 18.6% and 'Happy' at 8.4%. Thus, 80% of the emotional expression is taken by fear and sadness, indicating that, relatively speaking, the prevailing tone of the video was negative. With almost just as little, there was also a marked absence of surprise (1.8%) and anger (4.2%), suggesting that the emotional atmosphere of the video sustained, rather than shifted, in its intensity.

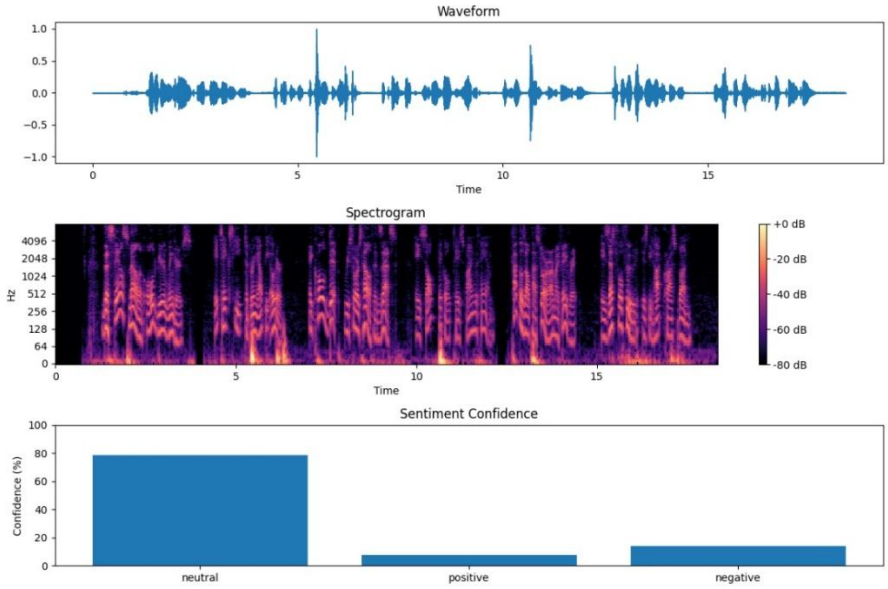

Fig.7. Audio sentiment analysis results showing the waveform (top), spectrogram (middle), and sentiment confidence levels (bottom) indicating a predominantly neutral emotional tone.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Fig.7. demarcates the outcome of a sentiment analysis conducted on an audio input by a multimodal emotion recognition model. The waveform is displayed at the top subplot and is a plot of the amplitudes of the audio signal over time. Peaks and troughs denote parts of speech activity, while sharp intensity spikes at certain time intervals denote parts of emotional articulation. The middle subplot is the spectrogram, with frequency (in Hz) plotted against time. It shows the time-frequency distribution of the sound signal with intensity of color employed to indicate energy content (dB). Active speech regions are marked with clear vertical bands and energy clusters, and brightness at lower frequencies indicates the presence of fundamental tones and harmonics characteristic of human emotions. The bottom subplot, titled Sentiment Confidence, indicates the output probabilities of three sentiments: neutral, positive, and negative. The model has a predominantly neutral tone with an estimated confidence of approximately 78%, with the positive and negative sentiments being much lower, with approximately 9% and 13%, respectively. The results indicate that the speaker had a steady and calm tone in the entire audio sample, with very little emotional change being identified by the model. This type of analysis is vital in the analysis of vocal affective states in tasks such as human-computer interaction and mental state assessment.

Late fusion gave a strong boost in performance for the system. The multimodal ensemble model attained 92% final accuracy, well above the individual modality performances. This means that merging cues from all three channels provides a better understanding of emotions than relying on any one modality alone. The system is also robust under real-world scenarios, including varying speech speeds, diverse facial expressions, and emotionally laden text.

5. CONCLUSION

It is comprehensive of emotion recognition in various modes, taking advantage in terms of the three modalities, i.e., text data, audio data, and visual data, to provide the exhaustive methodology in this paper. The combination of separate and state-of-the-art deep learning architecture into DistilBERT for text sentiment analysis, Vision Transformer (ViT) for an emotion-detecting view, and a specific convolutional neural network (CNN) for emotion classification using audio from RAVDESS dataset has produced a highly efficient multi-pronged system for perceiving human emotions.

On the other hand, all unimodal models performed well within their individual datasets; unsurprisingly, DistilBERT scored best, given that it captured the context of language very well. On the other hand, ViT topped the bunch in facial expression reading, while the enhanced CNN model excelled in captive registration of vocal emotions. But the core strength or value in this research primarily is in the multimodal fusion approach, in which the fusion of predictions of all three modalities dramatically increased the overall accuracy attained in emotion detection. This is the true power of using such multiple, complementary sources of data for attempting to tackle an issue that is quite delicate and complicated: understanding emotions.

The findings further support the prediction that in emotional intelligence-relevant activities, multimodal systems would be more advantageous compared to unimodal systems most especially for real-life applications such as human-computer interaction, virtual assistants, and mental health monitoring systems. Future works may include real-time implementation of emotion detection and promising research in complex fusion methods such as attention-based or transformer-level multimodal feature fusion for better performance and contextual understanding.

REFERENCES:

[1] M. Pantic and L. J. M. Rothkrantz, “Toward an affect-sensitive multimodal human- computer interaction,” Proc. IEEE, vol. 91, no. 9, pp. 1370–1390, Sep. 2003.
[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.
[3] R. Cowie et al., “Emotion recognition in human-computer interaction,” IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32–80, Jan. 2001.
[4] P. Ekman and W. V. Friesen, “Facial action coding system,” Consulting Psychologists Press, 1978.

[5] D. Keltner and L. F. Barrett, "Emotion perception: A central feature of emotional experience," Emotion, vol. 1, no. 3, pp. 232–237, 2001.
[6] Y. Zhang, H. Wu, and L. Zhang, "Multimodal emotion recognition based on multi-feature fusion and decision-level fusion," IEEE Access, vol. 8, pp. 96540–96548, 2020.
[7] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, 2008.
[8] A. Vaswani et al., "Attention is all you need," Proc. NIPS, pp. 5998–6008, 2017.
[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019.
[10] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.
[11] A. Parikh et al., "A Decomposable Attention Model for Natural Language Inference," Proc. EMNLP, pp. 2249–2255, 2016.
[12] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," IEEE ICASSP, pp. 6645–6649, 2013.
[13] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," Multimedia Syst., vol. 16, no. 6, pp. 345–379, 2010.
[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Int. Conf. Learn. Representations (ICLR), 2021.
[15] R. Sanabria et al., "CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) Dataset," arXiv preprint arXiv:1803.05449, 2018.