

A Smart System for Recognizing Human Emotions based on Audio and Text in Multi-Faceted Environments

Dr. M.Sangeetha¹, Dr.K.Malarvizhi², J Clifford Lebo³, J Sakthivel.S.V⁴ and G Kirubakaran⁵

¹⁻⁵Coimbatore Institute of Technology, Coimbatore, India

Email: msangeetha@cit.edu.in, {malarvizhi@cit.edu.in, cliffordlebo02@gmail.com, sakthivelsivakumar2001@gmail.com, kirubakaran041@gmail.com}

Abstract— Recognizing emotions from speech and text data in multi-faceted environments is a challenging task due to the ambiguous nature of emotion itself. To improve the modeling of different aspects of conduct and capture emotional content in various styles of speaking, the project utilizes multiple deep learning models to extract robust features from speech data and various lexical and prosodic features from text data. To achieve robustness and context sensitivity, it employs multi-faceted networks that are insensitive to outliers. The evaluation of models is done using various metrics such as accuracy, F-score, precision, and recall. In addition to deep learning models and multi-faceted networks, it also explores ensemble methods that combine different models to improve the overall accuracy and robustness of the emotion recognition system. These ensemble methods include combinations of random forest, XGBoost, multilayer perceptron, naive Bayes, logistic regression, and long short-term memory (LSTM) models. By combining multiple models, the project can leverage the strengths of each model while mitigating their weaknesses. Despite the diverse datasets used in the project, lighter machine learning-based models trained on a few hand-crafted features can achieve comparable performance to state-of-the-art deep learning-based methods for emotion recognition. Overall, the project aims to advance the field of emotion recognition in multi-faceted environments by improving the modeling techniques, utilizing multi-faceted networks, and evaluating the models with various metrics.

I. INTRODUCTION

Recognizing human emotions based on audio and text in multi-faceted environments is an important task with various applications in fields such as healthcare, education, and entertainment. A state-of-the-art system for this task should be able to accurately detect and classify emotions from different modes of input, while also being robust to noise, variability in speech patterns, and contextual cues.

To achieve this, the system can leverage recent advancements in deep learning and natural language processing techniques, such as transformer models. These models can be trained on large, diverse datasets of audio and text inputs labeled with their corresponding emotions, allowing them to learn complex patterns and relationships between inputs and outputs.

In addition, the system can incorporate various features and representations, such as acoustic features, prosody, and syntactic structures, to capture different aspects of human communication and enhance the accuracy of emotion recognition.

recognition. Furthermore, it can use contextual information from the surrounding environment, such as speaker identity, gender, and age, to improve the robustness and adaptability of the system. Overall, a state-of-the-art system for recognizing human emotions based on audio and text in multi-faceted environments is able to integrate various deep learning and NLP techniques, along with contextual information and feature representations, to accurately and robustly classify emotions from diverse sources of input.

II. RELATED WORKS

The system described in [1] utilizes Intelligent systems working together with pitch augmentation and shift augmentation which are needed to create new audio segments and work with minimal user effort, relying heavily on voice input. By using the DAIC-WOZ database and extracting the voice using COVAREP toolkit. The SVM classifier is utilized in this approach to discern between healthy and sad states by examining voice data. Results from the classification techniques are compared. An Android Application named CureD is used for self evaluation. A rule constraint generator [2] is a novel cross-representation speech model, inspired by disentanglement representation learning, to perform emotion recognition on wav2vec 2.0 speech features. A CNN-based model to recognize emotions with Transformer-based models. Furthermore combining the speech based and text based results with a score fusion approach. The method is evaluated on the IEMOCAP dataset in a 4-class classification problem.

An improved probabilistic model [3] relies on fusing correlation features of speech-visual proposes a multi-modal emotion recognition method. The speech and visual features are extracted using a two-dimensional convolutional neural network (2D-CNN) and a three-dimensional convolutional neural network (3D-CNN). support vector machines (SVM) complete the classification of multi-modal speech and visual emotion recognition. Using RML, eINTERFACE05, and BAUM-1 s datasets promotes a high recognition rate than other state-of-the-art methods.

The state-of-art detection results is achieved in [4] utilizes robust automatic speech-emotional-speech recognition architectures are proposed based on hybrid convolutional neural networks (CNN) and feedforward deep neural networks such as BFN, CNA, and HBN. BFN combines bag-of-Audio-word (BoAW) and feedforward deep neural network, CNA is based on CNN, and HBN is hybrid architecture between BFN and CNA.

The model proposed in [7] is CNN and LSTM were utilized to learn speech emotion characteristics. Simultaneously, many small-scale kernel convolution blocks were developed to extract information about facial emotions. Finally, This Method combined speech and facial expression features using DNNs. The IEMOCAP dataset was used to validate the multimodal emotion recognition model. The proposed model's total recognition accuracy has increased by 10.05% and 11.27%, respectively, as compared to a single modal of speech and facial expression.

III. PROPOSED SYSTEM

A. System architecture

Fig. 1 illustrates the entire flow of the method, showcasing the systematic progression of the system from input feeding to the final emotion classification. The depicted methodology provides a comprehensive understanding of how the system classifies emotions, offering a clear view of the functionalities of all the modules incorporated in the proposed system.

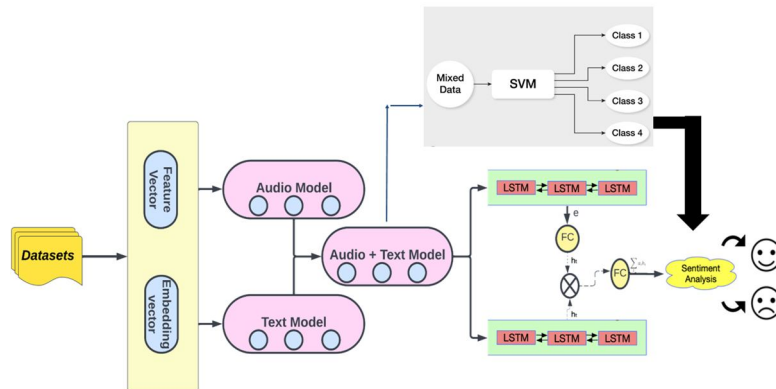


Fig. 1. Architecture diagram

The methodology begins with the collection of datasets, which are then subjected to a pre-processing stage to ensure the data's cleanliness and suitability for analysis. This involves various tasks such as noise removal, filtering, and normalization, aimed at enhancing the quality and standardization of the data. Once the pre-processing is complete, the data is transformed into feature and embedding vectors, which serve as the input for the subsequent audio and text modules. These modules consist of machine learning models that have been specifically trained on audio and text data, respectively. By leveraging the distinctive characteristics of each modality, these models generate predictions regarding the emotional state of the speaker or writer based on the provided input.

To further enhance the accuracy and reliability of the predictions, an ensemble learning approach is employed. This entails combining the individual predictions from the audio and text modules using appropriate ensemble techniques, such as voting or weighted averaging. By aggregating the outputs of multiple models, the ensemble learning approach aims to leverage the strengths of each modality and improve overall prediction performance. The integrated data, encompassing ensemble predictions, is then fed into a detection model, specifically an LSTM (Long Short-Term Memory) model. LSTM is a type of recurrent neural network that excels at capturing temporal dependencies and patterns in sequential data. By considering the temporal dynamics present in the emotional expression, the LSTM model generates a final prediction regarding the emotional state of the speaker.

To evaluate the performance of the model, various metrics such as accuracy, F1-score, and confusion matrixes are employed. These metrics provide quantitative measures of the model's effectiveness at accurately classifying emotions. Furthermore, the model's generalization and robustness are assessed by validating its performance across diverse environments and datasets. This ensures that the developed methodology can reliably handle real-world scenarios and effectively classify emotions across different contexts.

B. Dataset

To improve real-time prediction efficiency, data is gathered from multiple datasets including IEMOCAP, CMU-MOSEI, and other open-source datasets, creating a comprehensive Master Dataset. This diverse dataset enhances the model's capability to handle various emotional expressions and contextual variations.

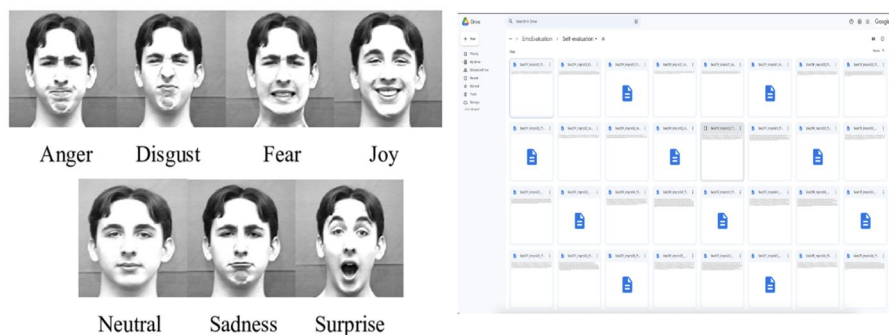


Fig. 2 Sample Dataset Images

Data collection is a systematic process of gathering and measuring information to address research questions and evaluate outcomes. In our study, we collected speech and text data from various sources to create a comprehensive Master Dataset. This dataset includes contributions from well-known sources such as IEMOCAP (Interactive Emotional Dyadic Motion Capture), which provides audiovisual data with speech, video, motion capture, and text transcriptions over a span of 12 hours. Additionally, we incorporated data from CMU-MOSEI, which offers 23,500 sentence utterance videos from diverse online YouTube speakers. The Master Dataset encompasses these datasets and others obtained from multiple sources, resulting in a diverse collection that represents seven different emotional classes. This comprehensive dataset enables more robust analysis and prediction of emotions in various contexts.

C. Data Preprocessing

Data preprocessing is an important step that involves cleaning and transforming the raw data into a format suitable for subsequent analysis. The pre-processing steps can vary depending on the nature of the data and the specific requirements of the project. For the audio data, pre-processing may involve filtering out noise and normalizing the amplitude to ensure consistency across the recordings. Feature extraction is then performed on the pre-processed audio data to extract relevant features such as pitch, energy, and spectral features. These features can be used as

input to machine learning or deep learning models. For the text data, pre-processing may involve tokenization, removing stop words, stemming or lemmatization, and converting the text into numerical representations such as Bag of Words (BoW) or TF-IDF.

Additionally, it is important to ensure that the labels of the data are consistent and accurate, and that the data is properly balanced across the different emotion classes to avoid bias in the models. Overall, data preprocessing is an important step in the above project to ensure that the data is ready for analysis and to improve the performance of machine learning or deep learning models.

D. Feature extraction and representation.

Feature extraction and representation are vital steps in analyzing audio and text data for emotion recognition tasks. When dealing with audio data, techniques such as Mel-Frequency Cepstral Coefficients (MFCC) are commonly employed to capture the spectral characteristics of the audio signal. Additionally, features like pitch, energy, and zero-crossing rate provide further insights into the audio's characteristics. These extracted features are combined into a feature vector representation, where each element represents a specific feature of the audio data.

In the case of text data, feature extraction techniques aim to convert textual information into numerical representations. The Bag of Words (BoW) approach treats the text as a collection of words and counts the frequency of each word, resulting in a numerical vector. Another technique, TF-IDF, not only considers word frequencies but also weighs the importance of each word based on its occurrence in the entire dataset. These feature vectors are standardized or normalized to ensure that all features contribute proportionally to the analysis, preventing any single feature from dominating the results.

Embedding vectors, on the other hand, play a crucial role in representing the semantic meaning and contextual relationships of words or phrases in natural language processing tasks. Techniques like Word2Vec or BERT map words to high-dimensional vector spaces, where similar words are located closer to each other. These embedding vectors capture the semantic relationships between words based on their co-occurrence patterns. Leveraging these embedding vectors enhances the understanding and processing of textual data by machine learning models, enabling tasks such as sentiment analysis and language translation to be performed with higher accuracy and efficiency.

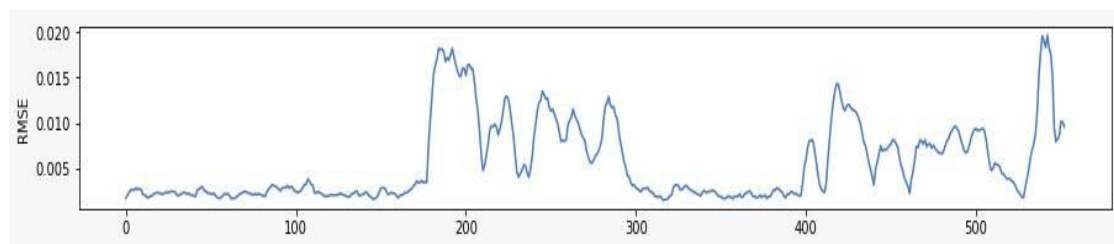


Figure 3: Feature vector

E. Learning Approaches.

1. Audio Processing Module.

The audio model could be implemented using a deep learning approach, such as a convolutional neural network (CNN) or a recurrent neural network (RNN) such as a Long Short-Term Memory (LSTM) network. The model would take the raw audio waveform as input and would use a series of layers to extract features from the audio data. These features could include Mel frequency central coefficients (MFCCs), which are commonly used in speech recognition tasks. The model would be trained on a labeled dataset of audio recordings, where each recording is labeled with the corresponding emotion. The model would be optimized to minimize the error between the predicted emotions and the ground-truth labels.

Once trained, the audio model would be integrated with the text model to produce a combined prediction of the emotional state of the speaker or subject. The combined prediction would be evaluated on a held-out test set to measure its performance and ensure its generalization and robustness across diverse environments and datasets.

2. Text Processing Module.

The text model would be responsible for analyzing and classifying the emotional content of written text, such as social media posts, emails, or chat transcripts. The text model could be implemented using a deep learning approach, such as a recurrent neural network (RNN) or a transformer-based model such as BERT. The model would take the raw text as input and would use a series of layers to extract features from the text data. These

features could include word embeddings, which represent each word in the text as a dense vector. The model would be trained on a labeled dataset of written text, where each text sample is labeled with the corresponding emotion. The model would be optimized to minimize the error between the predicted emotions and the ground-truth labels.

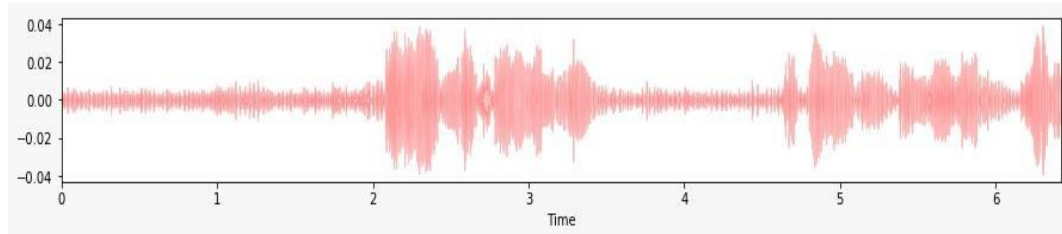


Fig 4: Audio Processing Module

3. Integrated Processing Module.

The fusion module would take the output of the audio model and the text model as input and combine them into a single prediction of the emotional state of the speaker or subject. The fusion module is implemented using weighted sum. For example, the output of the audio model and text model could be concatenated into a single feature vector and fed into a fully connected neural network for final classification. This model is a multimodal neural network, such as a multi-modal transformer or a multi-modal CNN-LSTM network. The performance of the fusion module would be evaluated on a held-out test set using appropriate metrics such as accuracy, F1-score, and confusion matrices. The module would also be tested on diverse environments and datasets to ensure its generalization and robustness.

4. Algorithms Embedded

RF (Random Forest) is a machine learning algorithm that builds a forest of decision trees and combines their outputs to make a final prediction. Each decision tree is trained on a subset of the data and features, making the algorithm robust to overfitting and able to handle high-dimensional data.

XGB (eXtreme Gradient Boosting) is a machine learning algorithm that uses gradient boosting to iteratively improve the accuracy of weak prediction models. It employs a novel technique to parallelize the tree construction and uses regularization to avoid overfitting.

SVM (Support Vector Machine) is a machine learning algorithm that finds a hyperplane in a high-dimensional space that maximally separates the data points into different classes. It works well for both linear and non-linearly separable data and can handle high-dimensional data with ease.

MNV (Multinomial Naive-Bayes) is a probabilistic machine learning algorithm that uses Bayes' theorem to make predictions. It assumes that the features are conditionally independent of each other given the class label and estimates the class probabilities using the frequency of the features in the training data.

LR (Logistic Regression) is a machine learning algorithm that uses a logistic function to model the probability of the binary outcome of a classification problem. It is simple and computationally efficient, making it a popular choice for classification problems with a large number of features.

F. *MLP (Multi-Layer Perceptron)* is a neural network architecture that consists of multiple layers of neurons, with each layer processing the output of the previous layer. It is a powerful machine learning algorithm capable of modeling complex nonlinear relationships between the input and output variables.

G. *LSTM (Long Short-Term Memory)* is a type of recurrent neural network that is capable of handling long sequences of data by remembering important information over long periods of time. It is often used for tasks such as text generation, speech recognition, and time series forecasting.

H. Algorithm of the Emotion Detection model

Input : Audio captured using a microphone.

Output : Identified Emotion of the speaker.

Begin

Step 1 : Initially a emotion detection module will be trained using an ensemble based machine learning module.

Step 2 : The module output will be stored in a pkl file for later detection purpose

Step 3 : The speaker's speech is recorded using a microphone.

Step 4 : The voice input will be inputted into the feature extraction module.

Step 5 : From the feature extraction module both audio and textual data will be extracted.

Step 6 : The extracted data will be loaded into the pre-trained ensemble based classification module which is done in step 2.

Step 7 : This module will return us the exact emotion of the speaker.
end.

I. Performance Evaluation

The method proposed in this work was evaluated and compared against state-of-the-art models using standard metrics based on the results. These metrics can be defined as:

Accuracy (eq. 1);

Recall (eq. 2);

Precision (eq. 3);

AUC - Area Under Curve;

F1-score (eq. 4);

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1\text{-score} = ((1+\beta^2) \frac{recall \times precision}{(\beta^2 \times precision) + recall}), \text{ with } \beta = 1 \quad (4)$$

TP stands for True Positives, TN for True Negatives, and FP for False Positives.

IV. IMPLEMENTATION

Ensemble Learning: Ensemble learning is implemented to improve the performance of the system. One way to apply ensemble learning is to train multiple models using different algorithms, such as Random Forest, SVM, XGBoost, Multinomial Naive-Bayes, and Logistic Regression. Then, the predictions of these models can be combined using a voting or averaging method to produce a final prediction. Another approach to ensemble learning is to use bagging or boosting techniques to create multiple variations of a single model. For example, a bagging technique like Random Forest can be used to train multiple decision trees on different subsets of the data, and their predictions can be combined using a voting or averaging method. A boosting technique like AdaBoost can be used to train multiple weaker models on different subsets of the data, and their predictions can be combined using a weighted sum. The performance of the ensemble models would be evaluated on a held-out test set using appropriate metrics such as accuracy, F1-score, and confusion matrices. The ensemble models would also be tested on diverse environments and datasets to ensure their generalization and robustness.

E1: Ensemble (RF + XGB + MLP)

The E1 ensemble model is a combination of three machine learning algorithms, including Random Forest, Extreme Gradient Boosting, and Multi-Layer Perceptron. By leveraging the complementary strengths of these algorithms, the E1 model can improve the accuracy and robustness of the emotion recognition system. The E1 model can be trained using various features extracted from audio and text data, including acoustic, lexical, and prosodic features. The model's performance can be evaluated using metrics such as accuracy, precision, recall, and F1 score. The Random Forest algorithm can handle noisy data and categorical data, while the Extreme Gradient Boosting algorithm can handle imbalanced datasets and large datasets. The Multi-Layer Perceptron can handle non-linear relationships and perform well on complex tasks.

E2: Ensemble (RF + XGB + MLP + MNB + LR+LSTM)

E2 is an ensemble model that combines six machine learning algorithms, including Random Forest, Extreme Gradient Boosting, Multi-Layer Perceptron, Multinomial Naive Bayes, Logistic Regression, and Long Short-Term Memory neural network. The LSTM neural network can capture the temporal dynamics of the input data and provide a more nuanced understanding of the emotional state of the speaker. The E2 model can be trained using various features extracted from audio and text data, including acoustic, lexical, and prosodic features, and can be evaluated using metrics such as accuracy, precision, recall, and F1 score. Overall, the E2 ensemble model with

RF, XGB, MLP, MNB, LR, and LSTM can provide a robust and accurate system for recognizing human emotions based on audio and text data in multi-faceted environment

IV. EXPERIMENTAL RESULTS

A. Consolidated results for audio only data.

TABLE I CONSOLIDATED RESULT FOR AUDIO ONLY DATA.

Models	Accuracy	F1	Precision	Recall
RF	76.0	77.0	78.2	78.8
XGB	75.6	76.0	77.9	78.8
SVM	77.7	78.2	79.4	80.5
MNB	75.3	76.9	78.2	79.3
LR	77.4	78.9	80.8	81.9
MLP	78.1	79.5	80.2	81.9
LSTM	77.6	78.4	79.2	80.6
ARE (4-class)	76.3	-	78.6	-
E1 (4-class)	86.2	85.9	87.6	87.9
E1	86.6	85.7	87.3	87.3

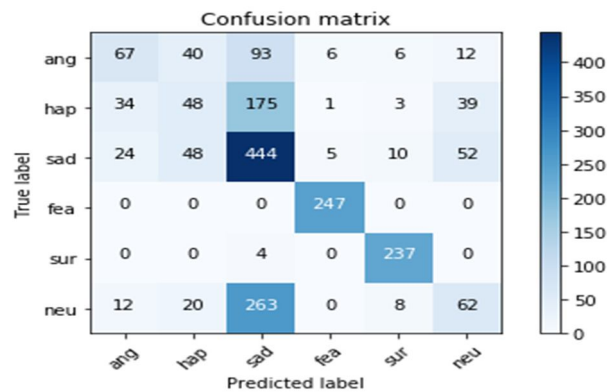


Fig 8 Confusion matrix for ensemble learning for audio only module.

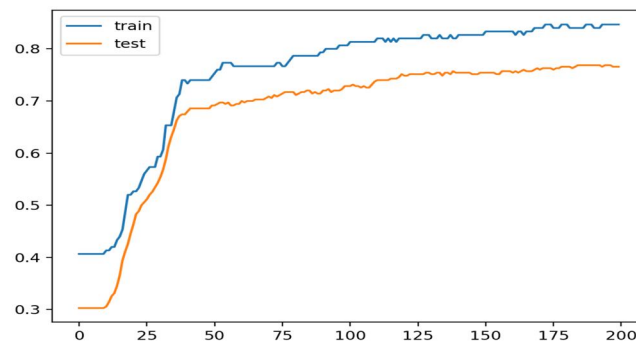


Fig 9 Epoch to accuracy graph for ensemble learning for audio only module.

B) Consolidated results for text only data.

TABLE II CONSOLIDATED RESULT FOR TEXT ONLY DATA.

Models	Accuracy	F1	Precision	Recall
RF	82.2	80.8	85.0	82.0
XGB	76.9	75.0	80.3	76.8
SVM	78.1	71.7	72.5	73.5
MNB	77.9	72.1	71.8	78.6
LR	74.2	74.3	69.5	77.3
MLP	76.6	75.5	72.4	75.0
LSTM	73.1	72.5	75.3	72.8
TRE (4-class)	75.5	-	73.5	-
E1 (4-class)	73.1	71.4	77.7	79.0
E2	74.9	76.0	81.4	73.2

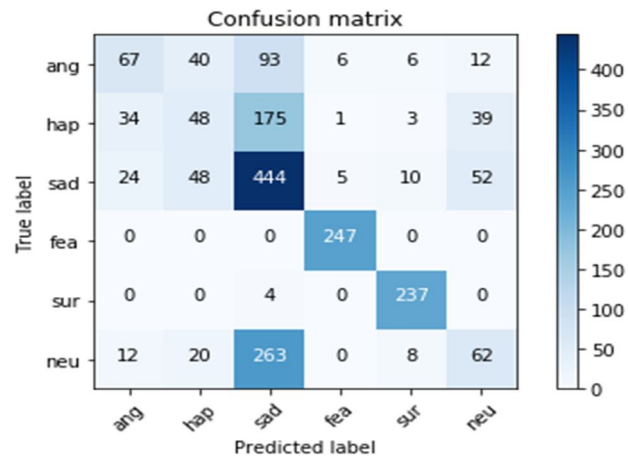


Fig 10 Confusion matrix for ensemble learning for text only module

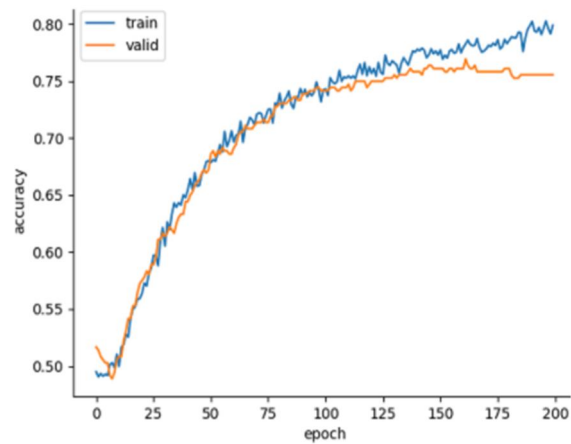


Fig 11 Epoch to accuracy graph for ensemble learning text only module.

C) Consolidated results for integrated dataset.

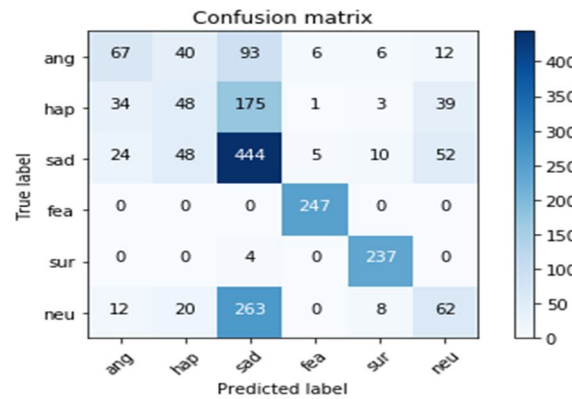


Fig 12 Confusion matrix for ensemble learning for the integrated module.

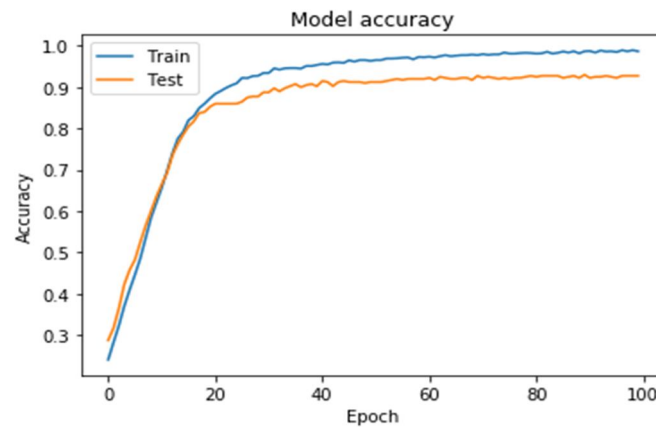


Fig 13 Epoch to accuracy graph for ensemble for the integrated module.

V.CONCLUSION

Emotional Instability is pretty unfortunate and should be considered a serious threat to the humans mental health and wellbeing. Recognizing human emotions based on audio and text in multi-faceted environments can have significant social impacts. By recognizing human emotions, mental health professionals can better diagnose and treat mental health disorders, such as depression and anxiety. Recognizing human emotions in real-time can help individuals better understand each other's emotions and intentions during conversations. This can improve communication and prevent misunderstandings in personal and professional relationships.

SCOPE FOR FUTURE

In future ,it would be interesting to delve in by analyzing the emotional content of online messages, social media platforms can detect and prevent cyberbullying, hate speech, and other forms of online harassment. Recognizing human emotions can help educators better understand the emotional state of their students, which can improve their teaching strategies and help them identify students who may need extra support. Overall, recognizing human emotions based on audio and text in multi-faceted environments has the potential to positively impact various aspects of society, from mental health to customer service to education.

REFERENCES

- [1] Fu, Yahui, et al. "Context-and Knowledge-Aware Graph Convolutional Network for Multimodal Emotion Recognition." *IEEE MultiMedia* 29.3 (2022): 91-100.
- [2] Chen, Luefeng, et al. "K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction." *IEEE Transactions on Industrial Electronics* 70.1 (2022): 1016-1024.

- [3] Bhavya, S., et al. "Speech Emotion Analysis Using Machine Learning for Depression Recognition: a Review." (2022).
- [4] Makiuchi, Mariana Rodrigues, Kuniaki Uto, and Koichi Shinoda. "Multimodal emotion recognition with high-level speech and text features." 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021.
- [5] Guanghui, Chen, and Zeng Xiaoping. "Multi-modal emotion recognition by fusing correlation features of speech-visual." IEEE Signal Processing Letters 28 (2021): 533-537.
- [6] Hsu, Jia-Hao, et al. "Speech emotion recognition considering nonverbal vocalization in affective conversations." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 1675-1686.
- [7] Hsu, Jia-Hao, et al. "Speech emotion recognition considering nonverbal vocalization in affective conversations." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 1675-1686.
- [8] Wani, Taiba Majid, et al. "A comprehensive review of speech emotion recognition systems." IEEE Access 9 (2021): 47795-47814.
- [9] Ezz-Eldin, Mai, et al. "Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition." IEEE Access 9 (2021): 19999-20011.
- [10] Hsu, Jia-Hao, et al. "Speech emotion recognition considering nonverbal vocalization in affective conversations." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 1675-1686.
- [11] Li, Mao, et al. "Contrastive unsupervised learning for speech emotion recognition." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
- [12] Cai, Linqin, Jiangong Dong, and Min Wei. "Multi-modal emotion recognition from speech and facial expression based on deep learning." 2020 Chinese Automation Congress (CAC). IEEE, 2019.
- [13] Zhang, Xue, Ming-Jiang Wang, and Xing-Da Guo. "Multi-modal emotion recognition based on deep learning in speech, video and text." 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP). IEEE, 2017.
- [14] Huang, Haiping, et al. "Multimodal emotion recognition based on ensemble convolutional neural network." IEEE Access 8 (2017): 3265-3271.
- [15] Huang, Jian, et al. "Multimodal transformer fusion for continuous emotion recognition." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.