IRJCS: International Research Journal of Computer Science
Volume 12, Issue 04, April 2025
ISSN: 2393-9842
https://www.irjcs.com/archives

Emotion Recognition System Using Machine Learning

Ruhi Shreya, Rohit Sharma, S Sudheshna
Vemana Institute of Technology
Banglore
Ruhishreya_2021@vemanait.edu.in, Rohitsharma_2021@vemanait.edu.in
ssudheshna@vemanait.edu.in

Publication History
Manuscript Reference: IRJCS/RS/Vol.12/Issue04/APCS10088 | Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.12/Issue04/APCS10088 Received: 06, April 2025, Revised: 12, April 2025 Accepted: 21 April 2025 Published Online: 05 May 2025 http://www.irjcs.com/volumes/Vol12/iss-04/08.APCS10088.pdf
Article Citation: Ruhi,Rohit,Sudheshna(2025). Emotion Recognition System Using Machine Learning. IRJCS: International Research Journal of Computer Science, Volume 12, Issue 04 of 2025 pages 160-164
doi:> https://doi.org/10.26562/irjcs.2025.v1204.08
BibTeX Ruhi@2025Emotion

Abstract: This paper introduces the concept of an "Emotion Recognition system using machine learning (ERS)" designed for recognizing and analysis of the emotions. The emotions that are detected are through speech, text or facial based data. It addresses key challenges, such as analyzing the patient's emotions through the facial expressions and the speech to monitor mental health, provide accurate feedback based on the text inputs and speech while using customer care feedbacks. The Emotion Recognition system using machine learning (ERS) uses different models for each data to provide accurate results. For the facial expressions sequential model is used, for speech Convolutional Neural Network (CNN) and Recurrent Neural Network(RNN) are used. The text data uses transformer-based models.

Keywords: emotion recognition; Convolutional Neural Network; Recurrent Neural Network; sequential; multimodal; machine learning; transformer-based models

1. INTRODUCTION

Emotion recognition is a crucial aspect in artificial intelligence (AI). This model is essential in present and future approaches. ERS can be utilized in Real time applications in various fields such as healthcare, customer services and intelligent AI assistant's, through this model the user experiences and decision-making increases significantly. This models are designed and implemented to ease the human work as, the huge amount of data which is collected and processed is given to the model for analysis. The analyzed data is then utilized to classify the real-time emotion that is shown by the human. A study in 2022 said that the early detection of anxiety and depression reduces the chances of affecting a person's mental well-being, through the continuous monitoring of patients diagnosis the improvement in mental health is verified. A deep learning model that uses speech and text for the emotion detection uses multimodal analysis to understand the human emotions. The features focus on the hidden emotions that add depth for the recognition[1]. This recognition model tracks emotions over time involving dynamic recognitions of emotions. To ensure the security of societies and other market places, this model is utilized in surveillance which detects suspicious behaviors in various high security zones. The advancements in the machine learning and artificial intelligence establish and involve human-AI interactions across multiple individuals and industries.

2.RELATED WORKS

An article suggested a new approach using both text and sound through a dual-modal framework to detect hidden emotions. They automated AED-FST dataset with a competing encoder-decoder framework and trained separate models for each modality: CNN and BiLSTM for audio, Text BERT and RoBERTa were separately used for the text. Their approach identifies hidden emotions by analyzing mismatches between audio and text model predictions [1]. The study also performed a detailed analysis of model architectures and features across several datasets like AED-FST and RAVDESS, showing BiLSTM's superiority over CNN in most cases. They also used hybrid feature sets, cross-modal fusion, and graph convolutional networks to discuss existing research gaps in emotion recognition systems in their literature review, to show emerging trends in emotion recognition systems[1]. A paper for facial emotion recognition using deep learing demonstrated a system capable of recognizing faces and emotions in real time utilizing CNN and dlib. The dataset fer-13 contains a grey scale images with processed pixel size. For face recognition, the system utilizes dlib face encoding which is then compared against real-time input. Emotion recognition is performed using a CNN model trained on FER-2013 dataset that utilized an input layer, the hidden layer and the output layer. It was noted that conventional methods, although swift and easy to understand, have fallen behind more complex approaches relying on deep learning due to the latter's performance in real-world settings. [3]. An article of hidden emotion detection framework that integrates speech and text data with deep learning developed a custom AED-FST corpus containing audio files and transcripts of question-and-answer dialogues that states that the speech would be translated into the text format for the data processing, each associated with both evoked and expressed emotions.

BiLSTM, CNN, BERT, and RoBERTa were among the models utilized across the various modalities. They proposed a new method for uncovering hidden emotions by evaluating the output discrepancies of parallel models which was achieved with a remarkable accuracy of 81.44% with MFCC-BiLSTM and RoBERTA fusion. They also drew upon classical methods such as Eigenfaces, LBP, and Fisherfaces which were beneficial in understanding the core concepts [5].

3.SYSTEM DESIGN

A. Design

The Emotion recognition system takes the data from the user in the form of text, speech or facial features that is taken for test analysis to predict the emotion behind the input. For the training of the models, preprocessed datasets are imported for each model of text, speech and face. The datasets for text include ISEAR, for facial expressions it is Fer13 and for the audio the datasets are obtained from RAVDESS. The datasets are trained from each models, For the facial expressions sequential model is used, for speech Convolutional Neural Network (CNN) and Recurrent Neural Network(RNN) are used. The text data uses transformer-based models. After the user provides the input, this goes to the test data. Based on the training data and its accuracy with respective to each emotion classifications, the output is delivered.

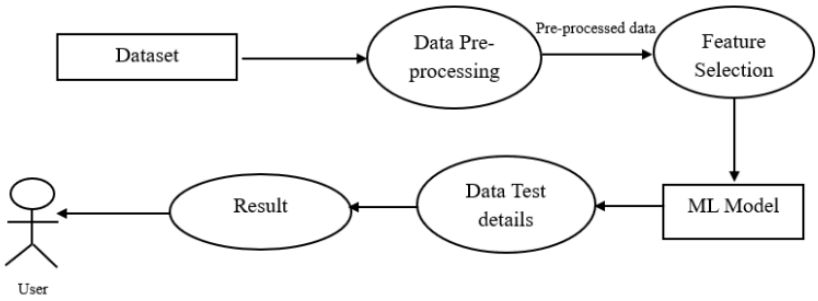

Fig.1. Data flow of ERS

The fig 1 represents the dataflow diagram of the emotion recognition model. First, the process initiates with the Dataset acquisition step where a Dataset is received. Subsequently, this is sent to the Data Pre-processing module, which is responsible for receiving data cleansing, normalization, and transformation, ensuring that all data is coherent and consistent. The data is them processed to the feature selection module which determines and selects the important relevant features to enhance the efficacy of the ML model accuracy and the efficiency of the ML model. The fed features are trained using the Machine Learning Model and then the results are subjected to prediction. The feature selection is based on selecting subset of wanted features from the whole dataset to reduce the model's complexity and also to improve the model's performance.

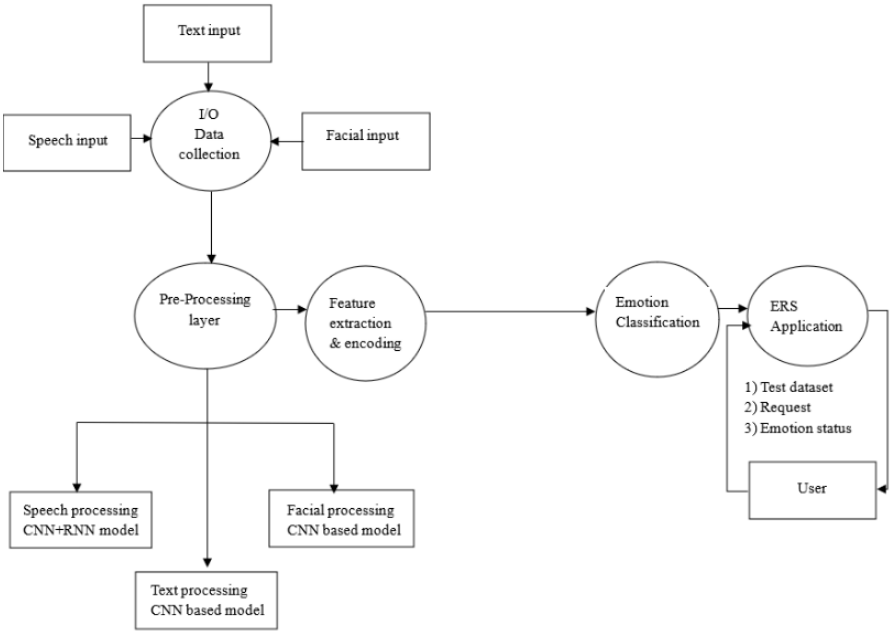

Fig.2. System architecture

The selected features are then pushed onto the machine learning models based on the input provided, which can be text, speech or facial expressions. These features are provided to the necessary models for the model analysis and its predictions. The result gets released and sent to the User providing the feedback and thus ending the interaction loop. The architecture is primarily focused on modularity whereby each stage can be optimized and updated in tune with the objectives of the specific stage independent from the others. This type of workflow would best work in customized regions such as the analysis and decision making in healthcare and finance and also in studying customer behavior where such decision focusing on data as the primary source is essential.

B. System architecture

The Fig.2 shows the system architecture that comprises a number of interrelated modules, each fulfilling particular tasks in an emotion recognition process. Starting from the I/O (Input/Output) Data Collection module, the system obtains information from three basic sources: text, speech, and facial data. These are then sent to the Pre-Processing Layer, wherein data is normalized, noise is removed, and the correct format is applied. Each modality is subsequently handled through deep learning techniques peculiar to each characteristic. During the preprocessing stage, the resulting data is sent to a Feature Extraction and Encoding module where important features pertaining to the emotion are extracted and encoded into a standard format, which in this case is the Unified Format. Subsequently, Emotion Classification is performed on the encoded features using pre-trained machine or deep learning models to classify emotions based on a set of defined classes. The emotion classification results are then fed into the ERS Application for user-level engagement and feedback operation, which can be evaluation using a test dataset, user request interaction, or showing current emotional state.

C. Methodology

The Emotion Recognition System (ERS) facilitates precise emotion recognition by employing a multimodal deep learning approach that processes text, image, and audio data simultaneously. The system uses three different datasets: ISEAR for classifying emotion-related texts, Fer-2013 for interpreting facial expressions, and RAVDESS for recognizing emotions in speech. To process this diverse data, ERS employs Convolutional Neural Networks (CNNs) for both text and video modalities. For text data, CNNs perform sentiment analysis through pattern recognition in word embedding matrices corresponding to different emotional states. Concerning video data, CNNs for facial feature extraction determine the presence of visual emotional cues. With this architecture, meaningful feature extraction from high-dimensional input data becomes efficient and scalable. For the facial data, Sequential model is applied to the fer-2013 dataset. The workflow of the data is in a straight line manner: The input layer, one or more hidden layers and the output layer. The output from the first input layer is given to the next layer that is the hidden layer, the output of this hidden layer is then passed onto to the output layers. For audio data, ERS employs a hybrid architecture comprising CNNs and Recurrent Neural Networks (RNNs). The CNN layers are tasked with extracting low-level acoustic features from raw speech signals, while the RNN layers, specifically Long Short-Term Memory (LSTM) units, learn the time-based dependencies and variability in speech attributes such as pitch, tone, and intensity. Subsequently, the outputs from all modalities are integrated through an ensemble classification approach to improve the model further.

4. EVALUATION AND RESULTS

The evaluation process of “Emotion recognition system” model is been tested against various datasets and machine earning models for each input types. This has been done to ensure that the model attains maximum accuracy also to classify emotions and predict the emotion classesfrom the input types. The evaluations for each of the training and the testing models are verified within the model predictions. The pre classified emotions from the datasets are compared to the test data outputs with the accuracy in predicted outputs; this states the percentage of original and predicted emotion classes.

A. RESULTS

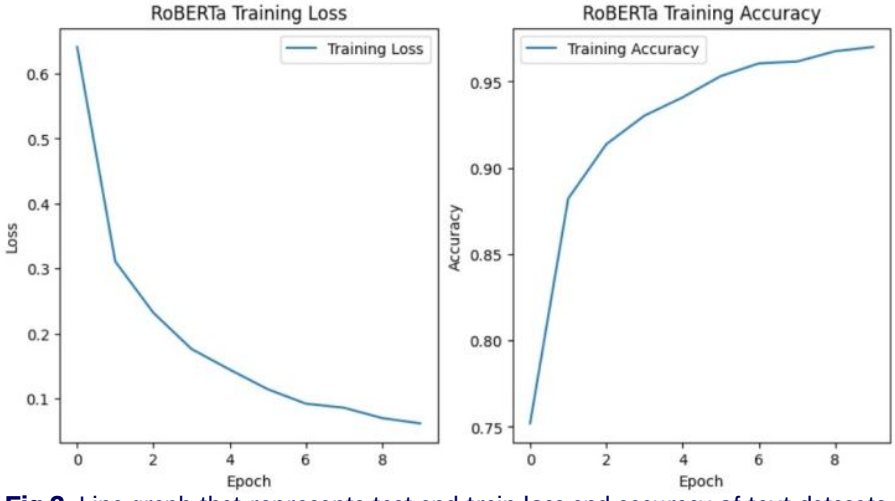

Fig.3. Line graph that represents test and train loss and accuracy of text datasets.

The Fig 3 represents the accuracy and loss of the text dataset model. The model used here is under CNN based model (RoBERTa). In the loss curve, there is a decline in training loss, starting above 0.6 and gradually decreasing to below 0.05. This indicates that the model is learning effectively by minimizing the prediction error during training. The initial accuracy for the first epoch was 75% and by the last epoch (9^{th})the accuracy increased to 97%.Hence, the model is perfect for the text classification.

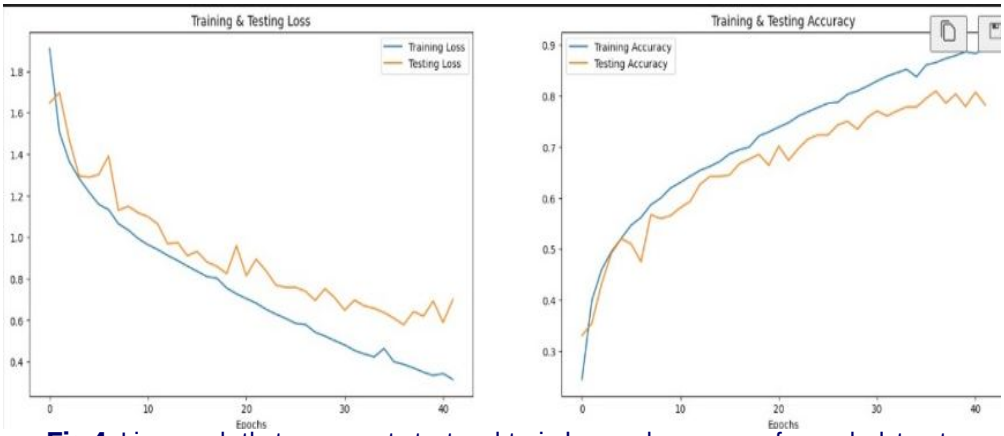

Fig.4. Line graph that represents test and train loss and accuracy of speech datasets

Fig 4 represents the line graph of training and testing performance of a hybrid CNN+RNN model over 45 epochs. The left graph shows the progression of training and testing loss, while the right graph shows the corresponding accuracy trends. The graph here shows the constant decrease in the test and the train loss by the CNN+RNN models to depict the audio data as an input. In the accuracy plot, the training accuracy steadily increases, eventually reaching 80.099%, while the testing accuracy also improves, attaining 80.9660%. The gap between the test and train accuracy shows that the model doesn't show overfitting.

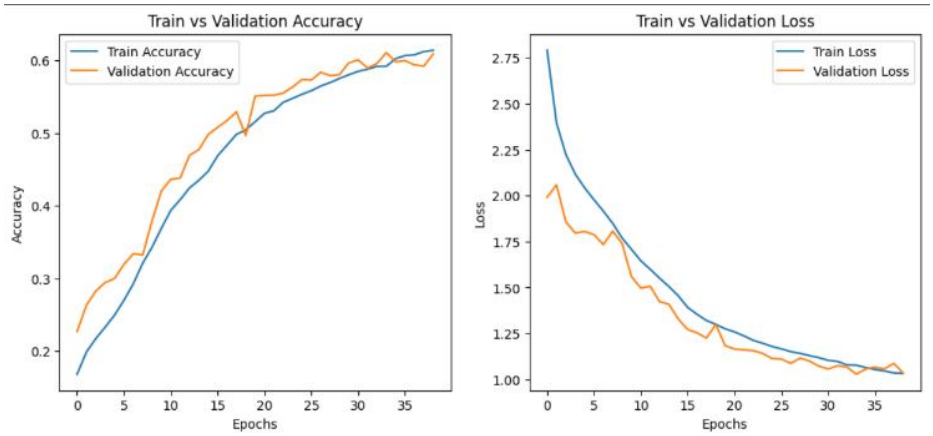

Fig.5. The training and validation accuracy and loss for facial dataset

Fig 5 shows the training and validation accuracy and loss for facial dataset Fer-13. The model applied was Sequential model and it achieved an accuracy of 60.14% and a loss of about 1.0458 on the test dataset. Similarly, the validation accuracy is 60.08% with a validation loss of1.0485**.** Since the training and validation accuracies are very close, and the losses are also aligned, this shows that the model is not overfitting. Other models were also built and considered for the emotion classification but, sequential model would perform better from the given datasets and acquired high accuracy when compared to the other models.

5. CONCLUSION

In order to improve the emotion recognition processes, the Emotion Recognition System combines text, speech, and facial analysis using deep learning models. Text emotion recognition uses RoBERTa-based transformer models, speech emotion recognition is executed with combined CNN and LSTM for temporal learning, and facial emotion recognition utilizes CNN models for spatial features extraction from images. With the integration of all three modalities, the understanding of emotional conditions is sophisticated, which enhances the system's agility towards complicated real-world stimuli. The system's outcomes undergo extensive testing, including unit, integration, performance, and end-to-end testing, guaranteeing defined and precise outputs. Enhanced generalization and stability of the model are achieved through advanced optimizers coupled with learning rate scheduling, data augmentation, and other domain-specific refined techniques. With all considerations taken, the multi-modal scheme provides state-of-the-art accuracy and robustness for emotion recognition and expands the applications for healthcare services, customer services, and human-computer interaction frameworks.

REFERENCES

1. R. T. Reddy and S. Palaniswamy, "Hidden Emotion Detection from Speech and Text Using Deep Learning Approach," in 2024 Second International Conference on Network, Multimedia and Information Technology (NMITCON), 2024, pp. 1-6. https://doi:/10.1109/NMITCON58196.2024.10553100
2. Sharma, V. Bajaj, and J. Arora, "Machine Learning Techniques for Real-Time Emotion Detection from Facial Expressions," in 2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON), 2023, pp. 1-6.
3. https://doi:10.1109/DELCON57910.2023.10127477
4. G. Singh, D. Singh, R. Sharma, and K. Bhardwaj, "Potential Approach for Text-Based Emotion Detection Using NLP Coupled With Deep Learning of Sentiment Analysis," in 2024 IEEE 15th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2024, pp. 1-5. https://doi:/10.1109/ICCCNT60824.2024.10553101
5. P. H. N. Ms, D. Ms, C. S. Ms, R. N. Ms, D. B. Mr, and S. H. Dr, "Face and Emotion Recognition in Real Time using Machine Learning," in 2022 Seventh International Conference on Communication and Electronics Systems (ICCES), 2022, pp. 1018-1025. https://doi:/10.1109/ICCES57224.2022.10553102
6. J. J. S and L. Jacob, "Multimodal Emotion Recognition Using Deep Learning Techniques," in 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2022, pp. 903-908.
7. https://doi:/10.1109/ICAC3N58196.2022.10553103
8. X. Zhang, M.-J. Wang, and X.-D. Guo, "Multi-modal Emotion Recognition Based on Deep Learning in Speech, Video and Text," in 2020 IEEE 5th International Conference on Signal and Image Processing, 2020, pp. 328-333.
9. M. deVelasco, R. Justo, J. Anton, M. Carriiero, and M. I. Torres, "Emotion Detection from Speech and Text," in Proceedings of the 2018 Workshop on Speech and Emotion, 2018, pp. 1-5. https://doi:/10.21421/MonSPEECH.2018-1-5
10. J. Kaur, Fahad, J. Saxena, S. P. Yadav, and J. Shah, "Facial Emotion Recognition," in 1st International Conference on Computational Intelligence and Sustainable Engineering Solution CISES-2022, 2022, pp. 528-533.
11. M. N. Alrasheedy, R. C. Muniyandi, and F. Fauzi, "Text-Based Emotion Detection and Applications: A Literature Review," arXiv preprint arXiv:2203.01253, 2022.
12. M. deVelasco, R. Justo, J. Anton, M. Carriiero, and M. I. Torres, "Emotion Detection from Speech and Text," in Proceedings of the 1st International Workshop on Multimodal Sentiment Analysis in Real-life Media, 2018, pp. 1-5.