IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A review of a system for sentiment analysis and emotion recognition based on deep learning that uses speech features and transcriptions

Sri Karun Maganti [(1)], Kande Vinay Harsha Vardhan[(2)]

Birla Institute of Technology and Science, Pilani – Hyderabad Campus, Hyderabad 2024

Abstract: Sentiment analysis and emotion recognition play a pivotal role in understanding human communication and interaction, with applications spanning from customer feedback analysis to human-computer interaction. In recent years, deep learning techniques have demonstrated remarkable success in these domains, particularly when incorporating speech features and transcriptions. This review delves into the state-of-the-art systems that employ deep learning for sentiment analysis and emotion recognition, focusing on those leveraging both speech features and transcriptions as complementary modalities.

The review begins by providing an overview of the theoretical foundations of sentiment analysis and emotion recognition and the evolution of deep learning in these fields. It highlights the significance of multimodal approaches that fuse speech features and transcriptions to improve accuracy and robustness.

The paper discusses various deep learning architectures, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and more recent models like transformer-based architectures, and their utilization in sentiment analysis and emotion recognition tasks. Special attention is given to pre-trained models and transfer learning techniques that have proven to be highly effective in reducing the need for large labeled datasets.

The importance of benchmark datasets and evaluation metrics for assessing the performance of sentiment analysis and emotion recognition models is addressed. The paper concludes with an exploration of potential future directions, including enhanced multimodal fusion techniques, cross-modal sentiment analysis, and the integration of these models into real-world applications. By reviewing and synthesizing the current state of deep learning-based sentiment analysis and emotion recognition systems using speech features and transcriptions, this paper aims to provide insights into the progress made, challenges faced, and the promising avenues for future research in these critical fields of natural language processing and affective computing.

Keywords: Feature selection, Speech emotion recognition, deep learning, deep neural network, deep Boltzmann machine, recurrent neural network, deep belief network, convolutional neural network

1. INTRODUCTION

In today's digital age, the analysis of human emotions and sentiments has gained significant importance across various domains, including customer service, mental health, market research, and human-computer interaction. Understanding the emotions and sentiments expressed in spoken language can provide valuable insights into the emotional states of individuals, enabling better decision-making and enhancing user experiences. This has led to the development of advanced technologies that combine deep learning and natural language processing techniques to create powerful sentiment analysis and emotion recognition systems.

Our proposed system, titled "Deep Learning-Based System for Sentiment Analysis and Emotion Recognition Using Speech Features and Transcriptions," represents a cutting-edge solution designed to extract and interpret emotional content from both the acoustic features of speech and its transcribed text. This system addresses the growing demand for automated emotion recognition and sentiment analysis in a variety of applications, including but not limited to call center quality assurance, mental health monitoring, and voice-controlled devices.

Key Components of the System:

1. Speech Feature Extraction: The system begins by capturing and extracting relevant acoustic features from the spoken language. These features encompass parameters such as pitch, intensity, formants, and more.

2. Deep Learning Models: Our system leverages state-of-the-art deep learning architectures, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models.

3. Emotion and Sentiment Analysis: The deep learning models process the acoustic features and transcriptions simultaneously, enabling the system to predict the emotional states and sentiments expressed in the speech. Emotion recognition may include

categories such as happiness, sadness, anger, and more, while sentiment analysis can identify sentiments like positive, negative, or neutral.

Our system represents an innovative approach to sentiment analysis and emotion recognition, fusing the power of deep learning with speech and text data to create a versatile tool that can revolutionize the way we understand and interact with emotions in spoken language.

2. LITERATURE SURVEY

The majority of NLP solutions, including voice-activated systems, etc., need speech as input. It is standard practice to use Automatic Speech Recognition (ASR) systems to convert this speech input to text before performing classification or other learning operations on the ASR text output. ASR corrects speaker-independent variations in voice transcriptions. ASR systems use probabilistic acoustic and linguistic models to provide outputs with high accuracy, but they also lose a substantial percentage of speech from various users, which results in information that signals emotion from speech. Due to this gap, research interest in Speech-based Emotion Recognition (SER) systems has increased over the past few years.

Three crucial factors for an effective SER system

1. Choosing a reliable database of emotional speech.
2. Recognize useful characteristics.
3. Create trustworthy classifiers using machine learning techniques.

The SER system's key problem is the extraction of emotional features. A typical SER system operates by extracting information from speech, including spectral, pitch frequency, formant, and energy-related data. These features are then classified to forecast different classes of emotion. Speech data that identifies emotions may be independent of the speaker or speaker-dependent. Some of the methods used in conventional classification jobs include Bayesian Network Model, Hidden Markov Model (HMM), Support Vector Machines (SVM), Gaussian Mixture Model (GMM), and Multi-Classifier Fusion. In this paper, we present a reliable method for classifying emotions from speech features and transcriptions. In order to increase the accuracy of emotion identification, the goal is to collect emotional traits utilizing speech data, combined with semantic information from text. We propose various deep network topologies to categorize emotion using text and audio characteristics. The following are the current work's main contributions:

3. PROPOSED METHOD

The machine learning (ML) model used by the spoken emotion recognition system works. With additional fine-tuning systems to ensure the model performs as intended, the phases of operation are identical to those of any other ML project. Data collection, which is of utmost importance, is the primary action. All the conclusions and decisions that a developed model will produce are supervised data, which the model being built will learn from the data provided to it. The secondary activity, referred to as feature engineering, is a synthesis of several machine learning tasks carried out over the collected data. The numerous data definition and data quality issues are addressed by these systems. An algorithm-based prototype is developed in the third stage, which is frequently investigated as the core of an ML project. This model use a machine learning (ML) algorithm to learn about the data and program itself to respond to any new data presented to it. Developers commonly repeat the steps of creating a model and estimating it in order to compare the performance of different methods. Results are measured to aid in selecting the ML algorithm best suited to the p Dataset.

The Toronto Emotional Speech Set (TESS), a data set of English language speech, was used. The phrases were recorded to represent the following seven emotions: joyful, sad, furious, disgusted, fear, surprise, and neutral state. This dataset consists of 200 target words said by two women, one younger and the other older. This dataset consists of 2800 files in total. 91 actors contributed 7,442 original footage to the data set known as CREMA-D. The actors in these clips, who ranged in age from 20 to 74 and represented a variety of races and ethnicities (African American, Asian, Caucasian, Hispanic, and Unspecified), included 48 men and 43 women. A selection of twelve sentences were read by the actors.

Six different emotions—anger, disgust, fear, happiness, neutrality, and sadness—as well as four different emotion levels—low, medium, high, and unspecified—were used to deliver the sentences. There are 24 professional actors in the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset (12 females and 12 males), each of whom speaks two lexically related lines with a neutral North American accent. Speech expressions might be peaceful, joyful, sad, furious, afraid, surprised, or disgusted.

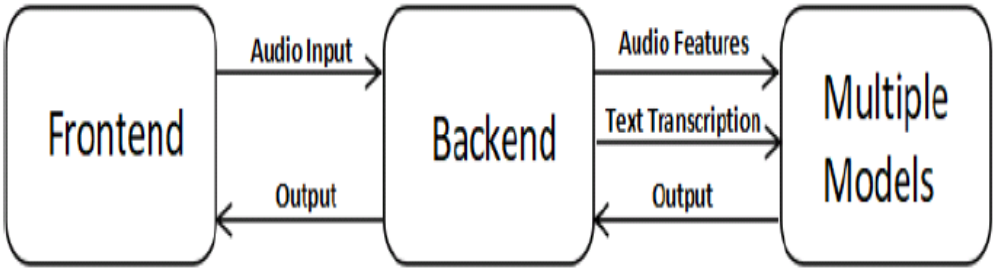


Figure 1: Block Diagram

4. DATA SET & DATA VISUALIZATION

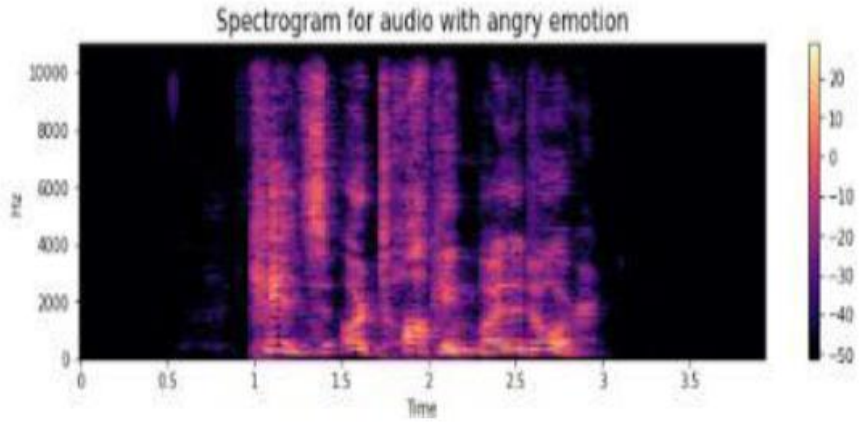
With regard to the file format and empty file input, the audio file used in this speech emotion recognition project is taken from the TESS Dataset and will be uploaded in.wav file format before the file upload process is validated. The audio file will then be connected directly to python files where the output is generated in the form of emotional labels. The given audio data in the document is described using data visualization.

We used the IEMOCAP dataset (Interactive Emotional Dyadic Motion Capture dataset) that researchers at the University of Southern California (USC) made available for this operation. This dataset includes text transcriptions for five sessions of conversations with ten speakers that total close to 12 hours of audio-visual recording. Two actors converse with one another for roughly two hours during each session. Eight classified emotional descriptors, including Anger, Happiness, Sadness, Neutral, Surprise, Fear, Frustration, and Excited, are used to categorize it. The dataset is already broken down into audio files at the sentence level, so we have files with just one sentence in them. As a result, there are roughly 10,000 audio files in total, from which features are subsequently extracted.

Class	Count
Angry	860
Happy	1309
Sad	2327
Fear	1007
Surprise	949
Neutral	1385
Total	7837

Figure 2: Sample Count in Dataset

Note: Not all of the information is included in the table above because several audio recordings' emotions were marked as Others or XXX. which are meaningless and which also undergo some pre-processing. Visual and graphic format. Here, the initial dataset's emotional labels are separated out, and the entire set of data is then represented as a spectrogram graph and a wave plot diagram. Because a sign's spectrum of frequencies changes over time, a spectrumogram might serve as a visual depiction of that spectrum. An amplitude vs. time waveform is plotted using a wave-plot, where the primary axis is amplitude and the secondary axis is time.

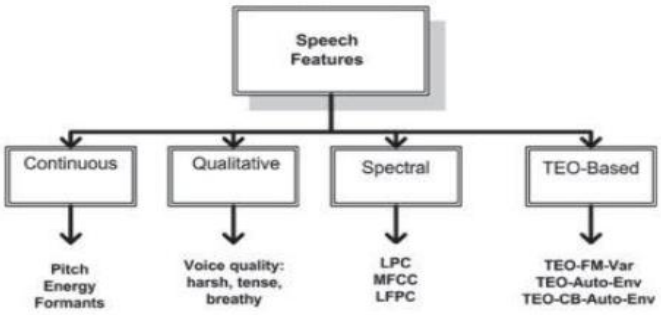


4.2 Features Selection:

These features—which extract the essence of an audio file—include Zero Crossing Rate, Chroma Shift, Root Mean Square Value, Mel Spectrogram, and MFCC (Mel information retrieval in the music category). When the frequency cepstral coefficient significantly shifts from positive to zero, this is when the zero- crossing rate occurs. These are a few of the most frequently applied audio features for audio content that conveys emotions, acoustic recognition, and to negative or from negative to 0 to positive. Mel Spectrograms are spectrograms that display audio on the Mel scale as opposed to the frequency domain. A signal's frequency must undergo a logarithmic change to create the Mel Scale.

4.3 Feature extraction:

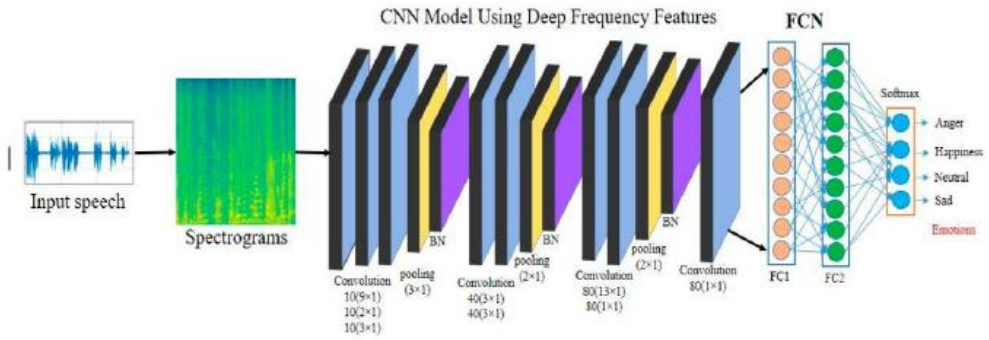
A significant amount of the voice signal's properties reflect emotional traits. What features should be employed is one of the difficult problems in emotion recognition. Recent studies have extracted a number of common features, including energy, pitch, formant, and some spectrum features, including linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC), and modulation spectral features. MFCC and modulation spectral features have been chosen in this study in order to extract emotional aspects.



The most popular way to characterize the spectral characteristics of voice signals is via the mel-frequency cepstrum coefficient (MFCC). These are the best since they take into account human perception sensitivity with regard to frequencies. The Fourier transform and the energy spectra were calculated for each frame and then transferred into the Mel-frequency scale. When speech signals are sampled at 16 KHz for our research, we extract the first 12 order of the MFCC coefficients. The mean, variance, standard deviation, kurtosis, and skewness are calculated for each order coefficient, and this is true for all other frames of an utterance. There are 60 dimensions in each MFCC feature vector.

4.4 Convolutional Neural Network

The speech data in the training set is trained using a convolutional neural network (CNN). The accuracy of the CNN model, RNN model, and MLP model training set and verification set is shown in Fig. 2. Figure 2 shows that the accuracy of the training set and verification set tends to rise with an increase in iteration times, particularly the training set. Finally, after 200 cycles, the accuracy of the training set is over 97% and that of the verification set is over 80%. In both the training set and the verification set, the CNN model created in this study outperformed RNN and MLP in terms of accuracy.



Convolution depicts the hierarchical extraction of voice characteristics, whereas maximal pooling removes superfluous data from the preceding layer features and streamlines the process. After every convolution layer, an activation layer is established, with the ideal activation function being discovered through trials. The output layer is configured with two completely linked neurons to categorize the speech signals into two groups. The network is further protected from over-fitting during training by random Dropout, which is set after each hidden layer due to the complexity of the network topology. After including Dropout, there will be a chance for each hidden layer's neurons to stop updating their weights during training, with a probability that is the same for all of them.

Algorithm

Step 1: As input, a sample of audio is given.

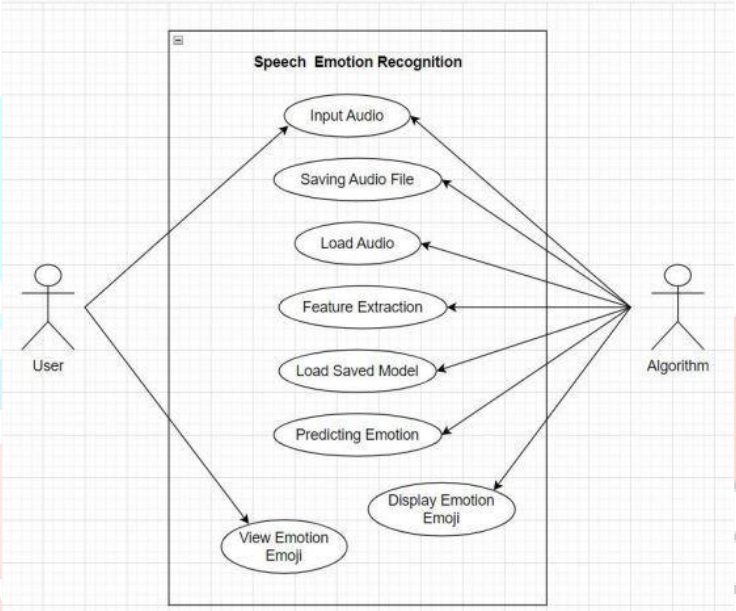
Step 2: The audio file is used to plot the Spectrogram and Waveform.

Step 3: We extract the MFCC (Mel Frequency Cepstral Coefficient) using the LIBROSA, a Python module, which is typically between 10 and 20.

Step 4: Remixing the data, separating it into train and test groups, and building a CNN model and its subsequent layers to train the dataset.

Step 5: Predicting human voice emotion using the training data (sample no. - predicted value - real value)

UML DIAGRAM*:*



5. IMPLEMENTATION

This section of the paper describes the various steps involved in the development of this project and the models incorporated.

5.1. Data Pre-processing

5.1.1. Audio

We discovered that the data is significantly imbalanced and contains labels that are inappropriate for inclusion in training data from our initial study and knowledge of the data. Then, because "happy" was underrepresented and "excited" was overrepresented, we blended examples from the two classes. Additionally, we eliminate examples marked as "others" or "xxx" because these labels refer to feelings that even humans cannot understand. These procedures are applied on the dataset, yielding a total of 7837 samples. Fig. 2 presents the final outcomes.

5.1.2. Text

Special characters were first deleted, then the text transcription was changed to lower case. **5.2. Feature Extraction**

We must first extract the characteristics from sound files and text transcriptions before we can extract emotions from audio recordings. So, we'll talk about the procedure we used to do this task here.

5.2.1. Text Features

Machine learning algorithms learn from a preset set of features from the training data in order to produce results for the test data. The main issue in language processing research, however, is that machine learning algorithms cannot operate directly on the unprocessed text. Therefore, certain feature extraction algorithms are required to convert text into a matrix (or vector) of features. The following are some of the most popular methods for feature extraction:

- **Bag-of-Words**
- **TF-IDF (Term Frequency-Inverse Document Frequency)**

A statistical technique called TF-IDF assesses a word's relevance to each sample in a group of samples. A word's frequency in a sample and its inverse document frequency over a set of documents are multiplied in order to achieve this.

- **Term Frequency (TF)**
- **Inverse Document Frequency (IDF)**

Term Frequency (TF):- The frequency of a term over the entire manuscript is indicated by its term frequency. It could be an idea of the likelihood of running into a term in the text. It assesses the frequency with which a term appears in a review relative to the number of words in the review. Term Frequency's behavior is precisely described by Eq. 1, while log normalization is defined by Eq. 2.

$$Tf = \frac{No.of\ times\ word\ occurs\ in\ review}{Total\ no.of\ words\ in\ review}$$

..Eq. 01

Inverse Document Frequency(IDF): An indicator of a term's rarity or recurrence across all documents in the corpus is the inverse document frequency. It concentrates on words that have a high IDF score or, to put it another way, words that are uncommon and appear in a small number of papers overall. The total number of documents in the corpus is divided by the number of documents containing the word, and the logarithm of the overall term is then calculated to provide the IDF, which is a log normalized value.

$$Idf(d,D) = log\ \frac{|D|}{\{d \epsilon D: t \epsilon D\}}$$

..Eq. 3

Given that the ratio inside the IDF's log function (Eq. 3) must always be bigger than or equal to 1, the IDF value is greater than or equal to 0.The ratio inside the logarithm approaches 1, and the IDF is closer to 0, when a term appears in a lot of papers.

5.2.2. Audio Features

We utilized the Python module librosa for feature extraction from audio files. which allows us to extract a number of features.

i. Signal mean: The mean, denoted by the symbol (mu), is a statistical term for a signal's average value. It is calculated by adding up all the samples and dividing by the total number of samples.

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} x_i$$

Eq. 04

Letting the index, i, range from 0 to N-1 will allow you to add the value in the signal, x i. Finally, divide that result by N.

ii. Signal Standard Deviation: The standard deviation is the same as the average deviation, except power is averaged instead of amplitude. Before averaging, the deviations are separately squared to achieve this. In order to make up for the initial squaring, the square root is then taken.

$$\sigma^2 = \frac{1}{N-1} \sum_{i=0}^{N-1} (x_i - \mu)^2$$

..Eq. 05

iii. Root Mean Squared Value (RMS Value): The square root of the mean squared value is used to represent the root mean squared value (or RMS value) of a signal over a given interval.

$$\mathrm{RMS(s(t))} = \sqrt{\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} s^2(t)dt}$$

..Eq. 06

iv. Pitch: Pitch is a perceptual characteristic of sounds that enables their classification on a frequency-related scale, or more often, pitch is the ability to assess the "higher" and "lower" pitch of sounds in the sense associated with melodies.

$$y[n] = \begin{cases} y[n] - C_l, & \text{if } y[n] \geq C_l \\ 0, & \text{if } |y[n]| < C_l \\ y[n] + C_l, & \text{if } y[n] \leq C_l \end{cases}$$

..Eq. 07

v. Energy: The energy of a signal is equal to the signal's overall magnitude. That roughly describes how loud the signal is for audio signals. The definition of a signal's energy is

$$E = \sum_{i=0}^{n} (|x(n)|)^2$$

..Eq. 08

A signal's root-mean-square energy (RMSE) is described as

$$E_{rms}^2 = \frac{1}{n} \sum_{i=0}^{n} (|x(n)|)^2$$

..Eq. 09

5.3 Machine Learning Model

We'll discuss the many machine learning models that were employed and put to the test when this project was being developed in this part.

5.3.1. Logistic Regression Classifier:

Various observations are categorized using a data classification method called an LR classifier. Logistic regression excels in binary classification challenges due to its binary character.

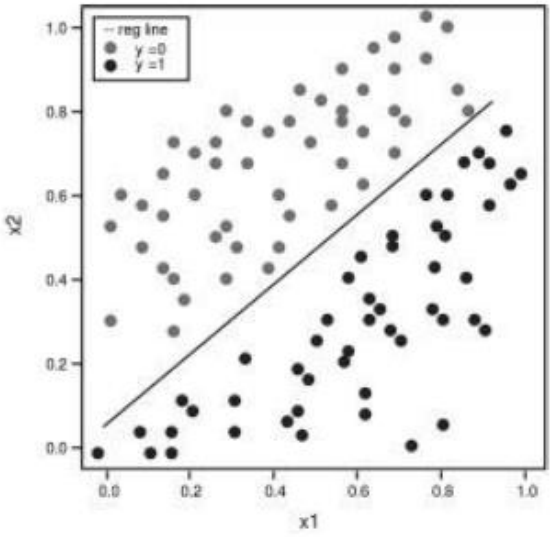


Fig. 3 - Binary LR Classifier Graph

So, it has been applied in a general sense here. There are six classifiers that have been applied to each emotion category.

5.3.2. Multinominal Naive Bayes Classifier:

The Naive Bayes family of algorithms uses the Bayes theorem (Eq. 10) to predict the category of a given sample under the assumption that each attribute is independent of the others. They will utilize the Bayes theorem to determine the probability for each category because they are probabilistic classifiers. The category with the greatest likelihood will be chosen. An illustration of this is the MNB Classifier, a Naive Bayes component used for multiclass classification.

$$P(A/B) = \frac{P(B|A) * P(A)}{P(B)}$$

..Eq. 10

5.3.3. Random Forest Classifier:

The RF Classifier is a classification model built on supervised learning and is an extension of the Decision Tree Classifier. They work by creating a number of decision trees during training and distributing the outcomes in accordance with the rules. As seen in Figure 4, the dataset is given to the tree. Each node in the tree represents a circumstance that dictates which branches it will give rise to. These circumstances form the basis for the anticipated output label.

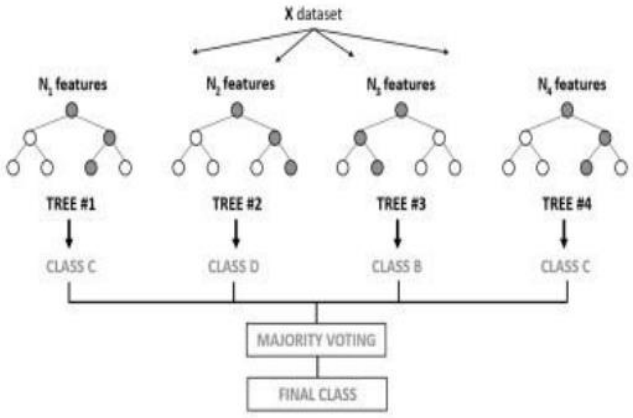


Fig. 4

5.3.4 K-Nearest Neighbour Classifier:

Both classification and regression use KNN, the most fundamental and significant method. The distance between each sample and each other piece of data is calculated by KNN. We categorize the new samples in accordance with this division, as seen in Fig. 5, which shows how a KNN model divides the sample into different classes.

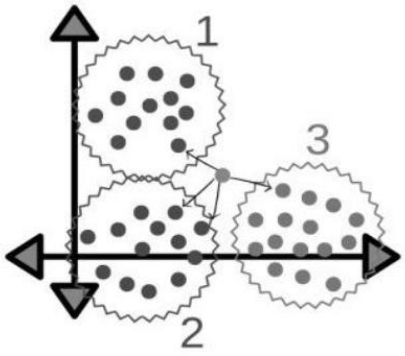


Fig. 5

The number of classes into which the data is to be divided, K, must be set at the proper value. Figure 6 illustrates how the error rate decreases as the value of K rises and then gradually starts to rise.

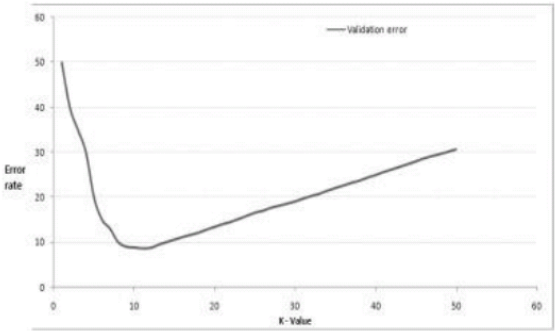


Fig. 6

5.3.5. Support Vector Machine:

A statistical learning model for classification is the SVM. The main goal is to use various kernel function types to project nonlinear separable samples onto higher dimensional space. The basic objective of this classifier is to draw a boundary or line dividing the various classes. These boundaries allow us to determine the classes of fresh datasets. In the training set, an ideal hyper plane is discovered, as shown in Fig. 7. Similar to this, several planes are set up for various courses.

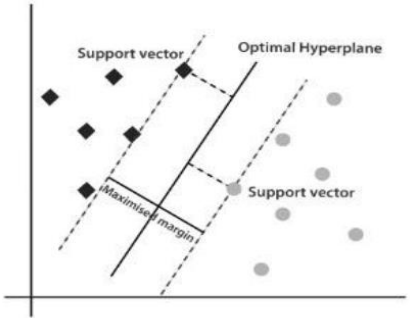


Fig. 7- 2 Class SVM Classifier

5.4 Deep Learning Model

In this section, we'll go over the many deep learning models that were tried out while this project was being developed.

5.4.1. Multi-Layer Perceptron: MLPs fall within the feed-forward neural network category. This model has at least three layers, namely an input layer, a hidden layer, and an output layer, as the name would imply. As we enlarge, the number of hidden layers increases, and the iterations rise, these models become considerably more refined and expressive. When the number of hidden layers and iterations go beyond a certain point, accuracy starts to decline. Figures 8 and 9 depict a single perceptron and various levels of interconnected perceptrons, respectively.

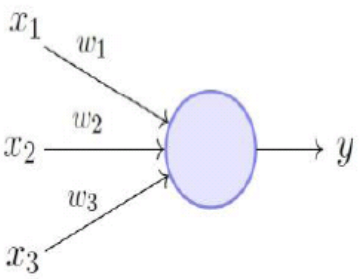


Fig. 8- Perceptron

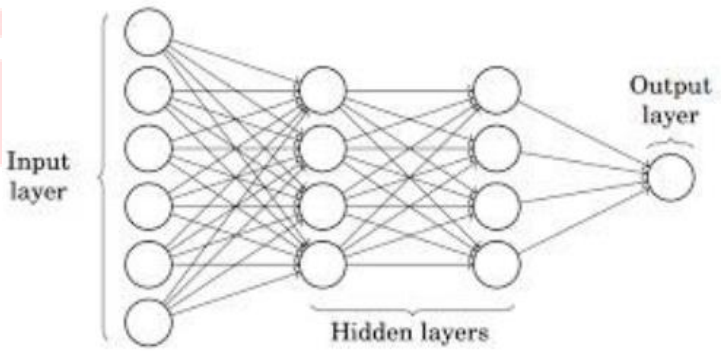


Fig. 9- Multi Layer Perceptron

5.4.2. Recurrent Neural Network:

Recurrent neural networks (RNNs) are a subset of artificial neural networks in which connections between nodes build a directed graph over time. It is permitted to display temporal dynamic behavior because of this.

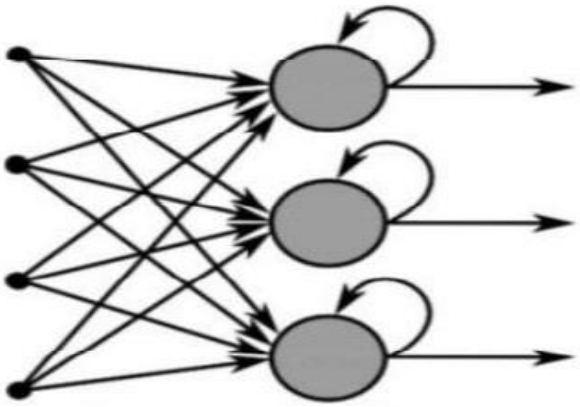


Fig. 10- Recurrent Neural Network

As illustrated in Figure 10, a hidden layer's output is transmitted back into the layer. resulting in better outcomes.

5.4.3. Long Short Term Memory:

Artificial Recurrent Neural Network architecture is what LSTM is. It can be applied to whole data sequences in addition to a single data point. In actuality, it performs better with longer data sequences. Since there can be lags of an arbitrary amount of time between significant occurrences in a time series, LSTM networks are suitable for categorizing, processing, and making predictions based on time series data.

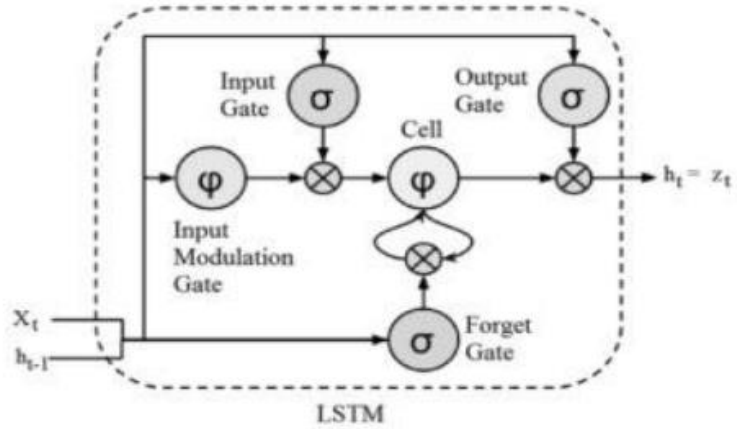


Fig.11-LSTM Cell

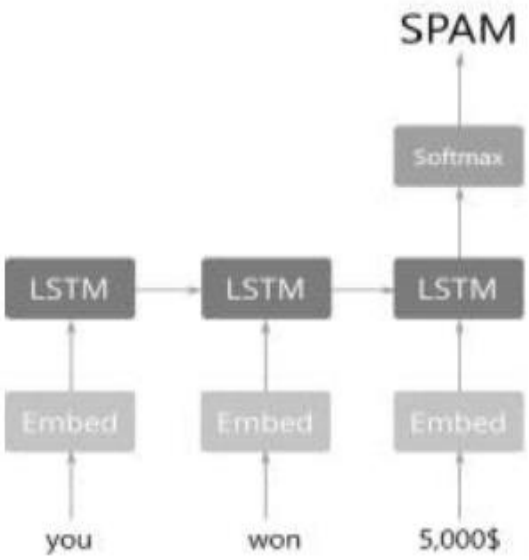


Fig.12-LSTM Classifier Sample

5.5 Model Evaluation

In this section, we'll go over the various evaluation techniques we used to examine the model results.

5.5.1. Accuracy: Sample percentage that is accurately classified.

$$Accuracy = \frac{Correct\ Prediction}{All\ Prediction}$$ ..Eq. 11

5.5.2. F-score: The F-score is a method of combining the model's recall and accuracy, and it is regarded as the harmonic mean of the model's recall and precision.

$$F\text{-}Score = \frac{T.P}{T.P + 0.5 * (F.P + F.N)}$$ ..Eq.12

5.5.3. Precision: Precision The percentage of relevant occurrences among the recovered examples is known as (also known as positive predictive value). Whether or not the measurements are precise, precision relates to how closely two or more measures are to one another.

$$Precision = \frac{T.P}{T.P + F.P}$$ ..Eq.13

5.5.4. Recall: The percentage of pertinent instances that were retrieved is known as recall. This metric reveals how many of the expected output's labels are accurate.

$$Recall = \frac{T.P}{T.P + F.N}$$ ..Eq.14

5.6 Front End (GUI)

The website we created for the front end, which users utilize to access the models, will be discussed in this part.

5.6.1. HTML5- The markup language known as HTML, or Hyper Text Markup Language, is used to organize and present content on websites.

5.6.2. CSS3- A style sheet language called CSS is used to describe how a document published in a markup language like html will appear when viewed.

5.6.3. JavaScript- A programming language that adheres to the ECMAScript specification is called JavaScript, or JS. A high-level programming language called JS is employed as the foundational technology in webpages to give them functionality.

5.6.4. FastAPI- Based on standard Python, FastAPI is a quick web framework for creating APIs with Python 3.6+. FastApi is built on open standards for APIs and is quick and very high performance, reliable, intuitive, and simple.

6. RESULTS

We'll talk about the outcomes of the various models from the previous section in this section. **6.1. Test Scores-**

6.1.1. LR Classifier-

Accuracy = 0.651
F-score = 0.658
Precision = 0.716
Recall = 0.634

6.1.2. MNB Classifier-

Accuracy = 0.564
F-score = 0.531
Precision = 0.760
Recall = 0.500

6.1.3. RF Classifier-

Accuracy = 0.697
F-score = 0.698
Precision = 0.743
Recall = 0.691

6.1.4. KNN Classifier-

Accuracy = 0.593
F-score = 0.580
Precision = 0.578
Recall = 0.623

6.1.5. SVC Classifier-

Accuracy = 0.671
F-score = 0.677
Precision = 0.686
Recall = 0.680

6.1.6. XGB Classifier-

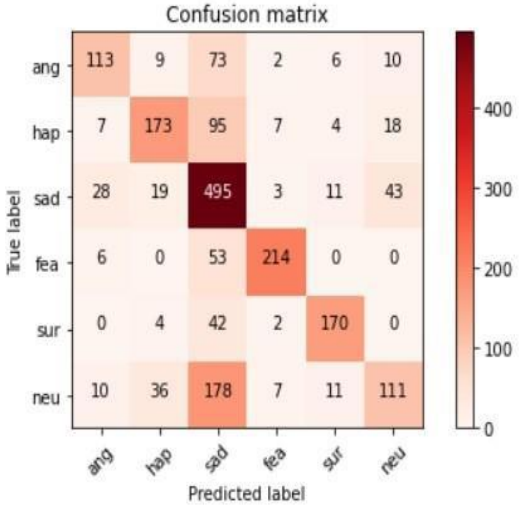
Accuracy = 0.596
F-score = 0.591
Precision = 0.681
Recall = 0.567

6.1.7. MLP Classifier

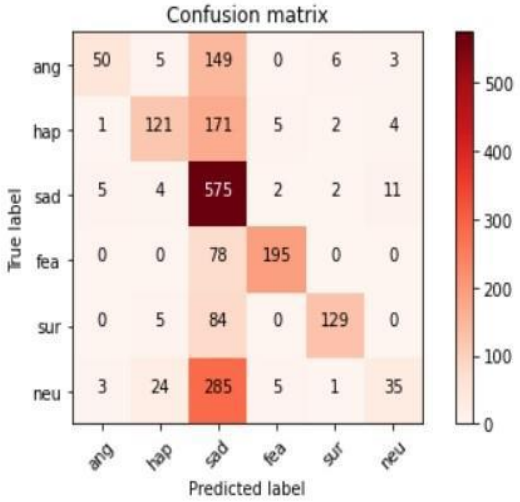
Accuracy = 0.684
F-score = 0.692
Precision = 0.702
Recall = 0.695

6.2 Confusion Matrix

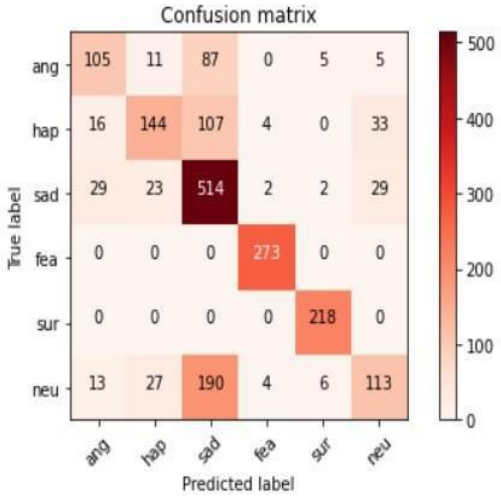
6.2.1. LR Classifier



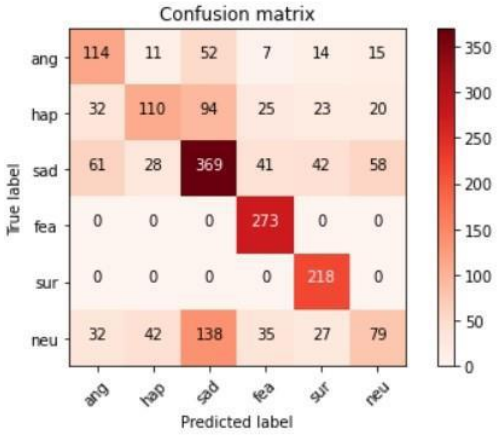
6.2.2. MNB Classifier-

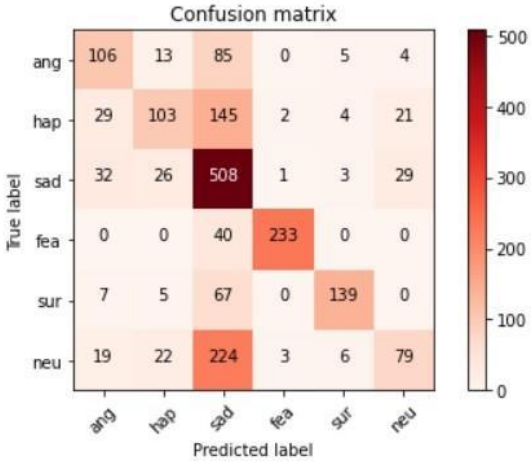


6.2.3. RF Classifier-

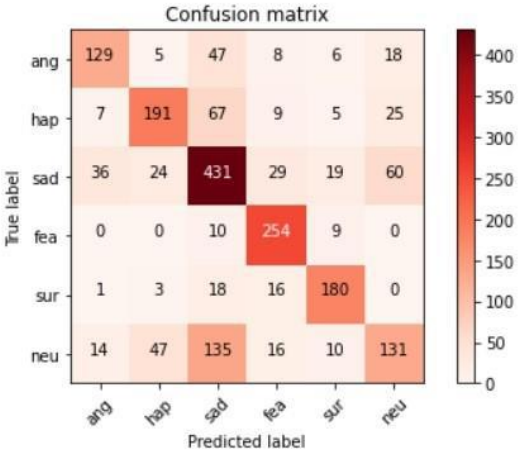


6.2.4. KNN Classifier-





SVC Classifier-

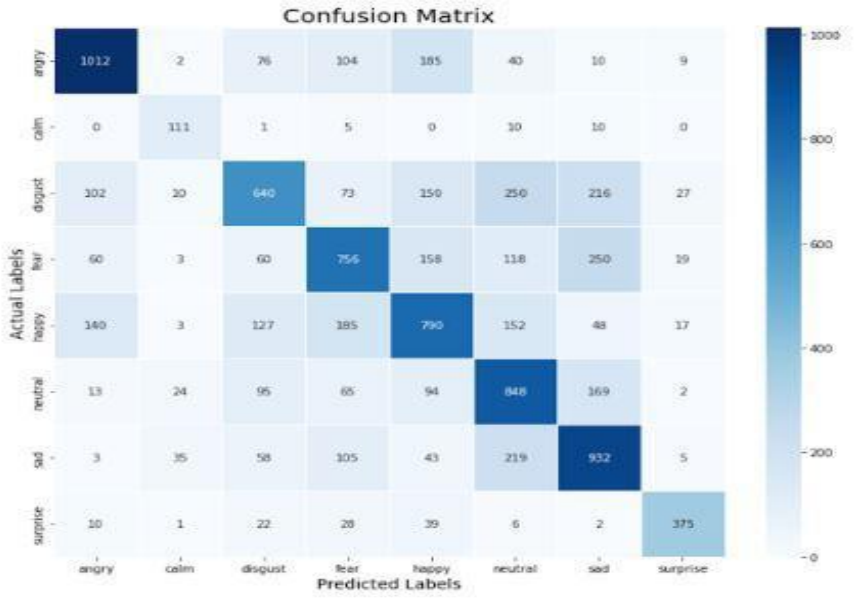


6.2.6. XGB Classifier

6.3 The average classification accuracy on RAVDESS is listed in Table

Following an experimental protocol, the derived models are tested in this paper's test set. SER's normalized confusion matrix is displayed in Table 1. Comparing the proposed method to previous works on SER, the new method achieves a higher recognition accuracy of expression recognition on the RAVDESS dataset. Table lists the RAVDESS's typical categorization accuracy.

	surprised	neural	calm	happy	sad	angry	fearful	disgust
surprised	0.88	0.02	0	0.04	0.06	0	0	0
neural	0.01	0.51	0.43	0.05	0	0	0	0
calm	0	0	0.96	0	0.04	0	0	0
happy	0.05	0.03	0	0.92	0	0	0	0
sad	0	0.15	0.05	0	0.75	0	0	0.05
angry	0.07	0	0	0.03	0.05	0.85	0	0
fearful	0.07	0	0	0.03	0.1	0	0.77	0.1
disgust	0	0	0	0	0	0.18	0	0.82



Perhaps surprisingly, the frequency of confusion between 'fear' and 'happiness' is as high as the frequency of confusion between 'sadness' or 'disgust'- perhaps this is because fear is a true multi-faceted emotion. Based on this, we will compare the characteristics of confounding emotions more carefully to see if there are any differences and how to capture them. For example, translating spoken words into text and training the network for multimodal prediction, and combining semantics for sentiment recognition.

7.CONCLUSION

We have carried out the task of emotion recognition in this work. We extracted characteristics from the Iemocap dataset, and then we used those features to extract emotions. Multiple ML and DL models were used, as previously mentioned. The models' accuracy score ranged from 0.6 to 0.7. Although the precision is poor, it is sufficient for the present and future work. Adding more audio features and employing more complicated models can greatly improve accuracy.

Even though it has been demonstrated that DL models do not outperform ML in terms of performance. However, more intricate and integrated models are anticipated to perform better. We have only integrated the audio and text aspects in this work. To get better outcomes, it would be best to use various fusion techniques. Observing how performance changes once these modifications are implemented will be fascinating.

In order to classify emotions into one of the eight categories, we built two concurrent convolutional neural networks (CNNs) to extract spatial features and a transform encoder network to extract temporal features. The TESS dataset under consideration is refined. We found it simple to categorize and feature this data because it contains noiseless data.By utilizing CNNs for sequence coding transformation and spatial feature representation, we were able to achieve an accuracy of 80.46% on the RAVDESS dataset's holdout test set. Based on a study of the findings, it is thought that speech to text conversion and semantic analysis could be used to recognize emotions.

In comparison to the current state-of-the-art techniques, the combined Spectrogram-MFCC model achieves an overall accuracy of 73.1% in emotion identification. When speech characteristics and speech transcriptions are combined, better results are seen. A class accuracy of 69.5% and an overall accuracy of 75.1% are provided by the combined Spectrogram-Text model, while a class accuracy of 69.5% and an overall accuracy of 76.1% are provided by the combined MFCC-Text model, respectively. These results represent improvements of 5.6% and almost 7% over current benchmarks. The suggested models can be applied to emotion-related applications, such as conversational and social robots, etc., where uncovering sentiment and emotion buried in speech may improve communication.

8. REFERENCES

[1]. Z. Huang, M. Dong, Q. Mao, and Y. Zhan, “Speech emotion recognition using cnn,” in Proceedings of the 22nd ACM international conference on Multimedia, pp.801–804,ACM.

[2]. S. Yoon, S. Byun, and K. Jung, “Multimodal speech emotion recognition using audio and text,” in 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 112–118, IEEE.

[3]. K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Fifteenth annual conference of the international speech communication association .

[4]. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” Language resources and evaluation, vol. 42, no. 4, p. 335.

[5]. A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mari˜no, “Speech emotion recognition using hidden markov models,” in Seventh European Conference on Speech Communication and Technology.

[6]. Surabhi V, Saurabh M. Speech emotion recognition: A review. International Research Journal of Engineering and Technology (IRJET). 2016;03:313-316

[7]. Nicholson, J., Takahashi, K. &Nakatsu, R. Emotion Recognition in Speech Using Neural Networks. NCA 9, 290–296(2000). https://doi.org/10.1007/s005210070006.
[8]. 'Han, Kun / Yu, Dong / Tashev, Ivan (2014): "Speech emotion recognition using deep neural network "
[9]. A. Rajasekhar, M. K. Hota, ―A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines‖, International Conference on Communication and Signal Processing, pp. 0114-0118, 2018.
[10]. Zheng, W. L., Zhu, J., Peng, Y.: EEG-based emotion classification using deep belief networks. In: IEEE International Conference on Multimedia & Expo, pp. 1-6 (2014).
[11]. Parthasarathy S, Tashev I.Convolutional neural network techniques for speech emotion recognition. In: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE 2018. pp. 121-125.
[12]. Matilda S. Emotion recognition: A survey. International Journal of Advanced Computer Research. 2015;3(1):14-19
[13]. Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. Asia- Pacific. 2017:1-4
[14]. G. Liu, W. He, B. Jin, ―Feature fusion of speech emotion recognition based on deep Learning‖, 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 193-197, 2018.
[15]. Koolagudi SG, Rao KS. Emotion recognition from speech: A review. International Journal of Speech Technology 2012.

Author Profile

SRI KARUN MAGANTI is a student currently in final year of M.Sc. (Mathematics), Birla Institute of Technology and Science, Pilani – Hyderabad Campus, Hyderabad

KANDE VINAY HARSHA VARDHAN is a student currently in B.E. (Hons) , Civil Engineering and Msc(Hons), Mathematics, Birla Institute of Technology and Science, Pilani – Hyderabad Campus, Hyderabad 2024