IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS ON VOICE DATA USING DEEP LEARNING

Deepa Yogish[1], Asha N [2], Yogish H K[3], Abhishek K L[4]

[1]Department of Information Science & Engineering, Don Bosco Institute of Technology, Bangalore 560074, India
[2]Department of Master of Computer Applications, National Institute of Engineering, Mysuru, India
[3]Department of Information Science and Engineering, Ramaiah Institute of Technology, Bangalore 560054, India
[4]Department of Master of Computer Applications, Ramaiah Institute of Technology, Bangalore 560054, India

Abstract : Sentiment Analysis (SA) is an application of Natural Language Processing (NLP) that is used to identify and extract sensitive information, such as a person's perspective on a particular piece of text. SA's main idea is to divide the authors' ideas on a variety of topics into positive, negative, or neutral categories. It works in a variety of domains including Business Intelligence (BI), politics, social science, etc. In recent years we have seen an increase in social networking websites, microblogs, and Web applications with a significant increase in user-generated data for sentiment mining. Data from online posts, tweets, videos, etc., all specific reviews on diverse subjects and events, provide big possibilities to have a look at and examine human opinions and sentiment. In this paper, we study how the sentiment of a voice data of a human can be classified into sentiment i.e., emotions possessed by the user while speaking. Emotions divided by the model are neutral, calm, happy, sad, angry, nervous, disgusted, and surprised. Our model captured and analyzed users' emotions and used them to improve user interaction / experience. The models for emotional recognition suggested here are based on the Deep Learning (DL) and Convolution Neural Networks (CNN). The main idea is to consider the Mel-Frequency Cepstral Coefficients (MFCC) normally referred to as the "spectrum of a spectrum", to be the only feature used for training the model.

IndexTerms **- Deep learning, Machine learning, MFCC, Convolution Neural Network, classifications, RAVDESS, TESS, Emotion detection**

1. INTRODUCTION

Human communication via audible communication is the basis for the exchange of information. It finds application in various fields such as calls centers and BPOs to get useful experience to express customer satisfaction about sales, strengthen communication, resolve ambiguities in language and computer programming compatibility according to the condition and feeling of the person.

In the proposed models the purpose is to detect only a sensor that contains a large number within the sound track. The CNN method is used to have a sensor that separates emotions such as computer vision or text calculations. In proposed work, main objective is to apply pure audio data using MFCC [4].

Structure of the paper

Section 2 describes about the literature survey in speech emotion recognition. Section 2 tells us about implementation details which includes datasets, algorithm and methodology. The proposed system architecture is depicted in Section 4 which describes about the architecture of the proposed system. Section 5 explain about the results discussion and section 6 concludes the paper.

2. Existing Work

In this area of research, a various number of classification algorithms are developed over the years. One such classification algorithm, proposed by Iqbal et al. [1] used Gradient Boosting, KNN, and SVM to work out granular separation in the Ryerson Audio-Visual Emotional Speech and Song (RAVDESS) database to identify the resulting differences on gender with absolute accuracy of 40% to 80%. Three kinds of datasets applied such as male, female, and mixed recordings. The 100% recognition for anger and neutrality for male recordings is achieved by RAVDESS, SVM and KNN. But the exciting and sad Gradient Boosting has done better than SVM and KNN. For only male recordings in RAVDESS, SVM achieves 100% accuracy with the same anger as half the male. The performance of KNN was also positive in neutrality with an accuracy of 87% and anger with an accuracy of 100%. KNN performance has been found to be worse in happiness and sadness compared to other traveler's groups. At the combined levels of male and female data, the performance observed by SVM and KNN was better with anger and neutrality than Gradient Boosting. CNN performance was really bad for joy and sadness. The Classifiers performance in only male recordings is better than only female recordings without SVM. For the available datasets, SVM gains more accuracy to only male dataset. Another approach presented by Jannat et al [2] obtained accuracy of 66.41% in audio data and accuracy of 90% in the mixed data set. In particular,

when looking at pre-processed image data involving faces and sound waves, the authors trained with three different networks: one net only works on image data, one on static audio waveforms, and a third on both data and waveform data taken together.
One of the earliest methods was to use the RAVDESS database, but it isolated it only from other obtainable emotions [8]. There are three types of song-sharing algorithms were suggested: first, a simple model was built, single workspace model and a multi-tasking model were developed. A single independent domain classifier was used by a simple model. Two different types of position sequences were used by the domain during training. The single-task model was used to train different stages of each domain. The multi-sectoral model has trained a multi-sectoral partner to share feelings which are shared across both the domains. Further during the test phase, the test data was categorized based on the predicted domain. standard classifier was used for data analysis to matches the estimated domain. Work has been done on the adoption of SVM-directed acyclic graph [7].

3. Methodology

The emotional recognition classifier models proposed here are based on CNN and DL strategy. The main idea is to consider the MFCC [4] which referred as "spectrum of a spectrum", being the factor used in model training. The MFCC is a different version of the Mel-frequency cestrum (MFC), and stands in its place in the field of sound recognition in automated speech recognition functions [5]. MFC coefficients were used as outcome of their ability to consider the amplitude spectrum of the sound wave in a normal vectorial manner. As introduced in [4], the audio file is split into frames of fixed window size to obtain statistically stable waves. The amplitude spectrum is usually reduced to a "Mel" scale. This function is performed with a sense of frequency and more beneficially in the vital reconstruction of the visible wave system of the human audible system. Around 40 features are extracted for each audio file. Features are framed by converting each audio file into a floating time series. Then, the MFCC sequence was formed from a time series. The same members of the MFCC are then transformed and the arithmetic mean is calculated in its horizontal axis.

3.1 Algorithms

The main reason behind CNN is to classify tasks as shown in Figure 1. The network can chip away at vectors with 40 features for every audio file given as input. These 40 features show the conservative mathematical type of the sound frame of length 2s. Subsequently, we give many training sets of size (40 x 1) on which one round of a 1D CNN is performed using a ReLU actuation work with a dropout of 20%, and a max-pooling function 2 x 2. The ReLU can be represented as g(z) = max {0, z}, and it permits us to acquire a huge value if there should be an occurrence of actuation via applying this function as a suitable decision to address hidden values. Pooling for this situation can help the model to focus just on the essential qualities of each part of information. We run the manner all over again with the aid of changing the kernel size. in addition, we applied some other dropouts after which we flattened the output to make it compatible with the following layers. finally, we follow a Dense layer (completely linked layer) with a SoftMax characteristic, various the output length from 640 elements to 8 and estimating the opportunity distribution of each class nicely encoded (impartial=zero; Calm=1; glad=2; sad=3; indignant=4; apprehensive=5;Disgust=6;amazed=7;)

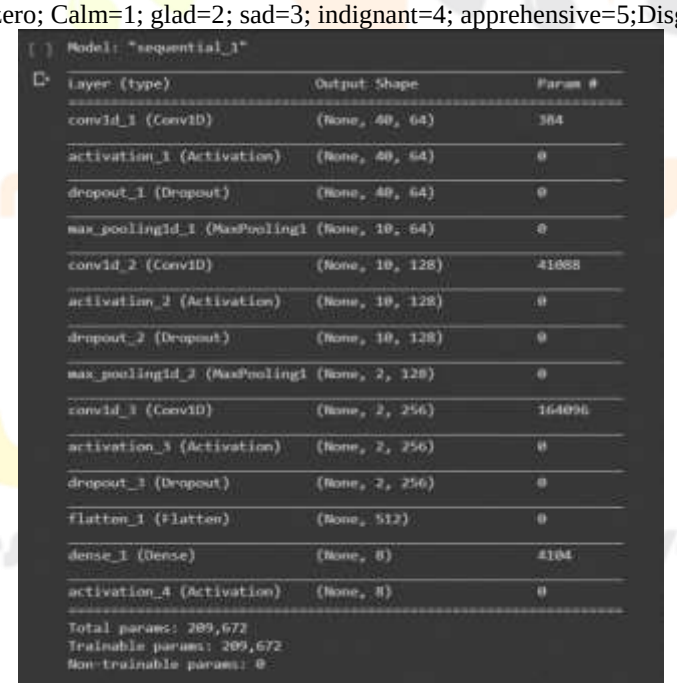

Fig. 1. Architecture of the proposed classifier

3.2 Datasets

The dataset used for our proposed work are RAVDESS) dataset [3] and Toronto Emotional Speech Set (TESS) dataset [8]. Samples of dataset contains 1440 speech files along with 1012 Song files from RAVDESS. It contains recordings from 24 professional actors include both female and male. Samples includes calm, happy, sad, angry, fearful, surprised, and disgusted expressions and the songs contain calm, happy, sad, angry, and fearful emotions.

4. Proposed methodology

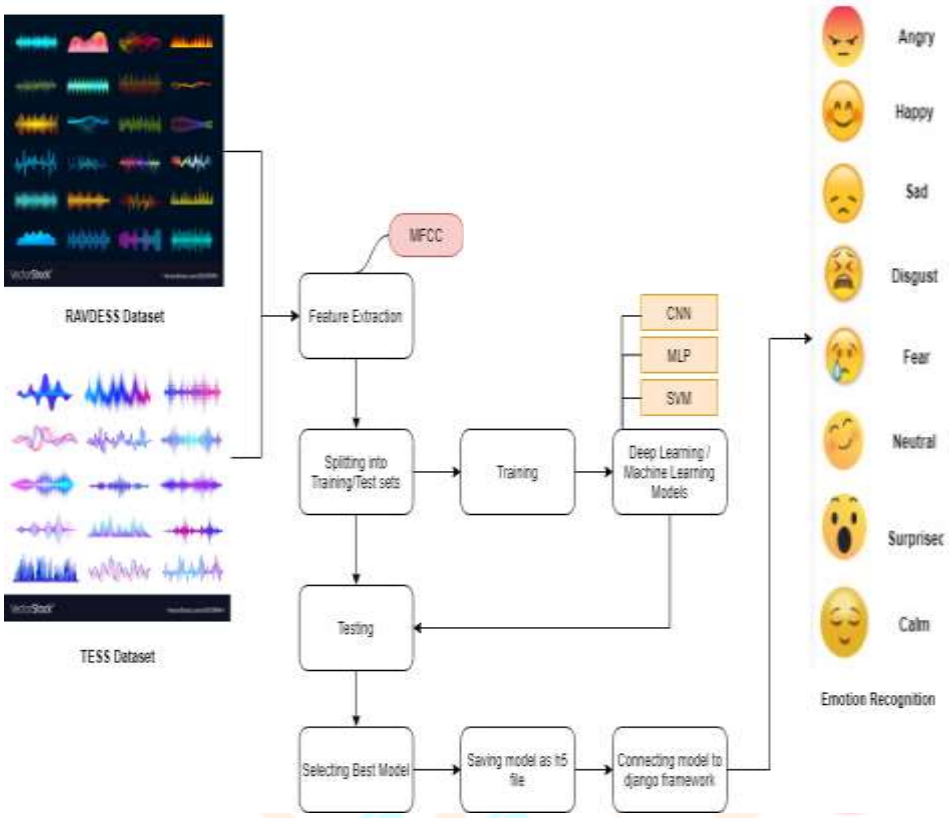

Fig.4.1 Proposed System Architecture

Figure 4.1 shows system architecture that defines the structure, behavior and more views of a system. Components of the system architecture are Feature Extraction, Training, Testing and Connecting model to Flask framework

For this we are going to make use of Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset and Toronto emotional speech set (TESS) dataset. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Librosa's load function will read in the path to an audio file, and return a tuple with two items. The first item is an audio time series (type: array) corresponding to audio track. The second item in the tuple is the sampling rate that was used to process the audio. The audio file is divided into frames, usually using a fixed window size, to obtain statistically stationary waves. We create our NumPy array for extracting Mel-frequency cepstral coefficients (MFCCs). Build a sequential CNN model with stacked Conv2D layers, activation functions, Maxpooling layers and Dense layers to perform classification. Developing a web application using Flask to give our project a user-interface to interact with the outside world.

5. Experimental Results

The results shown in Table 1 depicts the effectiveness of the model in the RAVDESS and TESS datasets. Table I shows the fit, recall, and F1 scores obtained for each emotion class. These results show that the fit and recall are very well balanced and the F1 scores for almost all classes are around the 0.85 value. The slight deviations in the F1 results indicate the robustness of the model. It is managed to effectively classify emotions into eight different classes. The "calmness" and "disgust" classes are less accurate models, but the introductory part observes not only the language, but also facial expressions and analyzes written text. Another representation of model reliability is shown in Figures 2 and 3. First, you can observe how the loss value decreases in both testing and training up to the 200th established epoch. The decline is less obvious from the 100th epoch, but it is still observable. Figure 3 shows the average accuracy for all classes. This increases with the number of epochs, as opposed to loss. It can be seen that these loss and accuracy values do not change much between training and testing. This will prevent the model from overfitting during training.

Table 1. Evaluation Metrics of the proposed model for each class

	precision	recall	f1-score	support
0	0.88	0.91	0.89	190
1	0.77	0.76	0.76	117
2	0.90	0.84	0.87	266
3	0.80	0.86	0.83	246
4	0.89	0.88	0.88	265
5	0.88	0.80	0.84	246
6	0.81	0.92	0.86	202
7	0.88	0.85	0.86	202
accuracy			0.85	1734
macro avg	0.85	0.85	0.85	1734
weighted avg	0.86	0.85	0.85	1734

Table 2. F1-Score for each class for proposed model

CLASS	**CNN**
SAD	0.80
ANGRY	0.89
HAPPY	0.90
DISGUST	0.81
SURPRISE	0.88
NEUTRAL	0.88
CALM	0.77
FEAR	0.88

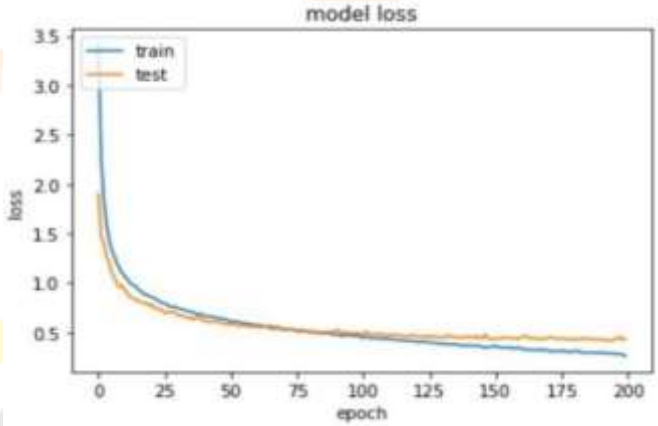

Fig. 2. Cost function of deep learning model for 200 Epochs

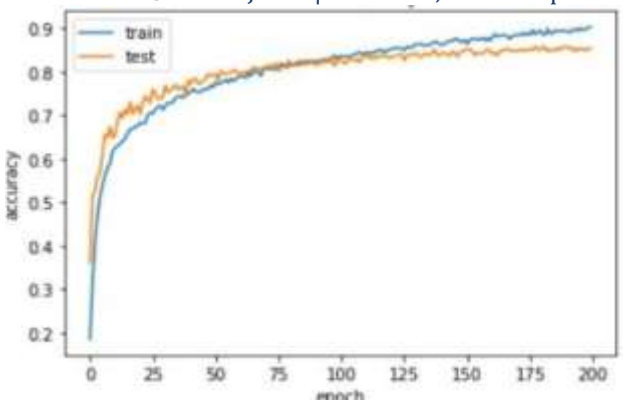

Fig. 3. Accuracy of deep learning model over 200 epochs.

6. CONCLUSION

Human communication is the basis for the exchange of information. It finds application in various fields such as calls centers and BPOs to get useful experience to express customer satisfaction about sales, strengthen communication, resolve ambiguities in language and computer programming compatibility according to the condition and feeling of the person. In the proposed models the purpose is to detect only a sensor that contains a large number within the sound track. The CNN method is used to have a sensor that separates emotions such as computer vision or text calculations. In proposed work, main objective is to apply pure audio data using MFCC. Our proposed work presented an architecture based on deep neural networks for the classification of emotions using audio recordings from the RAVDESS and TESS. The model is trained to classify seven different emotions (neutral, calm, happy, sad, angry, feared, disgusted, surprised) and 0.85 with the best and worst performance in the happy class (0, 90). Achieved an overall F1 score for the remaining classes (0.77). To achieve these results, we extracted the MFCC function from the audio file used for training. We then trained a deep neural network that uses 1D CNNs and Dense Layers to estimate the probability distribution of the annotated classes correctly. Our deep learning model obtained a F1 score of 0.86 on the test set. The performance achieved here suggests that such an approach based on deep neural networks is an excellent basis for solving the task.

References:

[1] Iqbal, A. and Barua, K. A real-time emotion recognition from speech using gradient boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–5.

[2] Jannat, R., Tynes, I., Lime, L. L., Adorno, J., and Canavan, S. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (2018), ACM, pp. 956–959.

[3] LIVINGSTONE, S. R. AND RUSSO, F. A.: The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English. PloS one 13, 5 (2018), e0196391.

[4] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

[5] Muda, L., Begam, M., and Elamvazuthi, I.: Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. arXiv preprint arXiv:1003.4083 (2010).

[6] Nair, V., and Hinton, G. E : Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (2010), pp. 807–814

[7] Platt, J. C., Cristianini, N. and Shawe-Taylor, J.: Large margin dags for multiclass classification. In Advances in Neural Information Processing Systems 12, S. A. Solla, T. K. Leen, and K. Muller, Eds. MIT Press, 2000, pp. 547–553

[8] Toronto emotional speech set (TESS) (https://tspace.library.utoronto.ca/handle/1807/24487).