Telegram Bot for Emotion Recognition Using Acoustic Cues and Prosody

Ishita Nag[1(✉)], Salman Azeez Syed[2], Shreya Basu[1], Suvra Shaw[3], and Barnali Gupta Banik[4]

[1] Department of Computer Science and Engineering, Institute of Engineering and Management, Kolkata, India
ishitanag04@gmail.com
[2] Department of Information Technology, Netaji Subhas University of Technology, New Delhi, India
[3] Department of Computer Science and Engineering, University of Engineering and Management, Kolkata, India
[4] C. R. Rao Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS), University of Hyderabad Campus, Hyderabad, Telangana 500 046, India

Abstract. In recent times, voice emotional recognition has been established as a significant field in Human Computer Interaction (HCI). Human emotions are intrinsically expressed by linguistic (verbal content) information and paralinguistic information such as tone, emotional state, expressions and gestures which form the relevant basis for emotional analysis. Accounting the user's affective state in HCI systems is essential to detect subtle user behaviour changes through which the computer can initiate interactions instead of simply responding to user commands. This paper aims to tackle existing problems in speech emotion recognition (SER) by taking into account the acoustic cues and prosodic parameters to detect user emotion. Here, the work will be mainly on the Ryerson Audio-Visual Database (RAVDESS) to extract Mel Frequency Cepstral Coefficients (MFCC) from signals to recognise emotion. The librosa library will be used for processing and extracting audio files before testing and classification is carried out using 4 different classifiers for a comparative study. This SER model is then integrated into a Telegram Bot to develop an intuitive, user-friendly interface as an application for psychological therapy.

Keywords: Speech Emotion Recognition (SER) · Multi-Level Perceptron (MLP) · Telegram bot

1 Introduction

1.1 Motivation

The introduction of the first modern computer system with a graphical user interface (GUI) in 1964 by Douglas Engelbert made the computer more accessible to the general public since it moved away from complex binary codes for operation, with this being

S. Mukhopadhyay et al. (Eds.): CICBA 2022, CCIS 1579, pp. 389–402, 2022.
https://doi.org/10.1007/978-3-031-10766-5_31

the first advancement towards human computer interaction (HCI). HCI later went on to become a significant research field in the 1980s. In HCI, the aim to improve the quality of communication between humans and computers is made much simpler and naturalistic through this GUI. Since then, there has been a paradigm shift towards multi-modal approaches where even voice interfaces are clubbed with GUI to develop adaptive, active and intelligent interfaces over passive and command-based ones which are much harder to operate. Such voice interfaces not only introduce speech recognition as in identification of words but also bring the possibility of understanding human emotions through various expressions. This gave rise to speech emotion recognition (SER).

Although there exist multiple social media platforms, Telegram, launched by brothers Nikolai and Pavel Durov, has around 400 million monthly active users with nearly 1.5 million new users joining each day making it one of the most popular messaging apps. A Telegram chatbot is an automated third-party application that runs within Telegram and can interact with users by sending messages. Telegram offers an API for developing bots for social interactions, CRM integration, ticketing systems, and other purposes. Other social messaging platforms, such as Facebook Messenger and WhatsApp, have strict chatbot rules, such as a 24-h window. However, there are no such rules in place with Telegram, which is a huge advantage. Also, unlike Facebook Messenger and WhatsApp, Telegram allows one to use chatbots in groups.

1.2 Aims and Objectives

This paper mainly focuses on machine learning techniques using these acoustic cues and its related parameters to drive emotion recognition through prosody and vocal bursts. SER is carried out over a large RAVDESS dataset, used to extract a broad range of 5 different features using 4 individual classifiers. A comparative analysis helps pick out the best result in accordance with accuracy, precision, weighted average, recall and f1-score. The second aim of this paper is to integrate the SER model into a Telegram bot. The Botfather bot and the python-telegram-bot library are the primary components required for our model's Telegram bot integration. BotFather is the default Telegram bot that assists in creating new bots and modifying existing ones. It generates the Telegram bot's Access Token, which is used during programming to interact with our functions. python-telegram-bot is an open-source wrapper python library for the Telegram Bot API which aids the development of various functionalities, primarily via the telegram.ext package. Joseph Weizenbaum developed the first NLP computer algorithm, ELIZA, in 1964. It was one of the first bot programs developed with the intention of simulating a psychotherapist. The first swell of AI technological innovation took the shape of chatbots. Voice-activated dialogic agents such as Siri, Google Now, Cortana, and Samsung S Voice are not classified as chatbots. Bots have progressed to serve a multitude of roles. Social media platforms enable developers to create a chatbot for their brand or service, allowing customers to carry out some of their daily activities from inside their texting platform. Telegram bots are relatively new in the chatbot space, but they have made a name for themselves.

2 Background & Literature Review

2.1 MFCC Extraction

Among the many SER approaches mentioned, most of these mainly target classification of MFCC features using a few different classifiers and their various combinations. There has been increasing focus over deep learning techniques due to their enhanced accuracy since they are able to learn complex acoustic and emotional cues and their ability to even capture low level features such as raw waveforms or spectrograms which simple classifiers cannot handle. The authors of [1] used such a deep learning method to learn emotion-related aspects domestically and globally from speech and log-mel spectrograms, using CNN-LSTM in both 1D and 2D. For speaker-dependent studies, the 2D CNN-LSTM achieved a precision of 95.33% on Berlin EmoDB and 89.16% on IEMOCAP. 3-stage support vector machine (SVM) has been utilized in [2] over Berlin EmoDB to classify 7 different emotions by extracting MFCC features to obtain accuracy of 68% whereas authors in [5] have implemented the KNN algorithm as classifier on Berlin EmoDB to extract MFCC feature to obtain 50% accuracy. Nalini et al. in [11] use AANN to extract spectral features such as MFCC and residual phase whose performance is evaluated using FAR, FRR and accuracy. Proves to show that combination of residual phase with MFCC can improve accuracy of SER.

2.2 Hybrid Classifier Models

Several papers also discuss newer hybrid methods in quest for increased accuracy and dataset versatility. Authors in [3] have employed HMM and SVM classifiers for the extracted MFCC and LFCC features from the SAVEE database. Four emotional areas are considered for emotion detection from speech, yielding a precision of 61.25% for MFCC and 46.88% for LFCC using the OAO approach of SVM on the gaussian kernel, which yielded the best results as compared to OAA and polynomial kernels. From this the authors has reached the conclusion that better emotional trends in speech data are observed using MFCC rather than LFCC and HMM classifier is better than SVM. Deep belief networks (DBN) have been employed in [6] instead of GMM to represent complicated and non-linear high-level associations between low-level features utilizing the HMM classifier. The hidden markov toolset (HTK) was used to extract MFCC from the benchmark FAU Aibo emotion dataset, resulting in a best UAR of 45.08% on 1-state HMM, compared to 44.1% and 43.5% on 3-state and 5-state HMM, respectively. Models that combine different HMMs have also been found to better represent emotion fluctuation, resulting in a UAR of 45.60%. Furthermore, in real-life circumstances, speaker nominalization provided the best UAR of 46.36%. Different noise models have also been employed in some articles to improve the quality of the speech at three SNR levels: 15 dB, 10 dB, and 5 dB. Because the database was collected in automobile settings, the roughly matched training and testing environment for LISA-AVDB could explain the good identification accuracy of loud speech at higher SNR. On EMO-DB, a combination of GMM as well as HMM based classifiers yield a recognition accuracy of 83.8%. Thus, it has been deduced that emotion identification in speech signals is strongly dependent on the acoustic environment [7].

2.3 Other Papers

Qiron Mao et al. in [10] focus on finding salient affect-related features on which the accuracy of SER depends. They use CNN for unsupervised, automatic feature learning from a few labelled samples and SVM classifier. SER accuracy is tested on 4 different public emotional databases to obtain robust performance in changing scenarios. Authors in [4] have extracted LPFC to represent signals and applied the HMM classifier for classification of 6 emotions and have achieved average and best accuracies of 78% and 96% respectively over a specifically designed emotion corpus.

Some papers have also shown the works done on gender and age classification. Hierarchical classification models were used for SER. Different feature sets were investigated based on the task of determining the speaker's age, gender, and emotion. The eGeMAPS dataset was critical in completing this assignment. Models were created using MLP neural networks. The findings revealed that having a distinct classifier for each gender and age group outperforms having a single model for both genders and ages [8].

3 Methodology

In this section, the proposed technique for the SER model has been discussed in detail along with each individual component involved such as the dataset used, classifiers applied and the features extracted.

3.1 Classifiers

For a more comprehensive comparative analysis, we use the same dataset over 4 different classifiers namely Multi-layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) classifiers. Each classifier differs in some aspects and when applied to different datasets, their performance also varies. Different classifiers also behave differently on the same dataset and since these classifiers or machine learning models are built based on certain assumptions on the characteristics of the dataset, that classifier performs the best whose assumptions are most in-line with the datasets. In the end, selection of any classifier depends on the use case. This is where a large dataset is useful to truly compare performances of varying classifiers.

3.1.1 Multi-Level Perceptron (MLP)

MLP is a kind of fully linked feedforward Artificial Neural Network (ANN), with only one hidden layer in the most basic version. MLPs have been proven to be capable of approximating an XOR operator as well as a variety of other non-linear functions. In supervised learning, MLP are frequently used and it uses backpropagation for training and stochastic gradient descent to optimize the log-loss function and as a result it trains iteratively as the loss function parameters are constantly updated. They learn to represent the correlation between inputs and outputs by training on a set of input-output pairings. The network can thus be seen simply as an input-output model, with the weights and biases serving as the model's free parameters [15]. The number of hidden layers and the number of units in these layers are important considerations in MLP architecture. Its

non-linear activation and multiple layers allow it to distinguish non-linearly separable data which is very useful for our SER analysis. We use the ReLU activation function which can be defined as:

$$f(x) = x + = \max(0, x)$$

where x denotes input to a neuron in the AAN. This activation function performs better gradient propagation to enable efficient computing in supervised learning models. The parameters that have been considered in the paper for building the MLP classifier are: alpha (value = 0.001), batch_size (value = 250), solver (taken as adam), epsilon (value = 1e−08), hidden_layer_sizes (value = (500,)), learning_rate (taken as adaptive), max_iter (value = 500).

3.1.2 Support Vector Machine (SVM)

SVM is a supervised learning model based on statistical learning frameworks following the structural risk minimization principle. Classification is done by constructing an N-dimensional hyperplane and using an efficient hyperplane searching technique which suitably separates input data to categories. Such searching helps reduce processing time by minimizing the training area. Training is done using this linear kernel function. Since it involves no assumptions on data type, it is able to circumvent overfitting. The parameters involved here are: kernel (taken as linear) and C, which is a regularization parameter (value = 1). SVMs identify the support vectors *vj*, weights *wfj*, and bias b to categorize the input information. For classification of the data, the following expression is used:

$$\mathrm{sk}(v, vj) = \left(\rho v^{e} vj + k\right)^{z}$$

In the equation mentioned above, k is a constant value, and b represents the degree of the polynomial. For a polynomial $\rho >$ zero:

$$v = ({}_{i=0}\Sigma^{n} w_{fj} \mathrm{sk}(vj, v) + b$$

In the equation mentioned above, sk represents the kernel function, v is the input, vj is the support vector, wfj is the weight, and b is the bias [16].

3.1.3 Decision Tree (DT)

These are non-parameterized supervised learning models which are given certain simple decision rules. Its' job is to use these decision rules to predict the values of target variables. This classifier can perform multi-class classification on datasets, which makes it ideal for SER utilizing a white box model. It also solves multi-output problems by building N independent models, one for each output given N inputs. The Decision Tree model's tree structure incorporates the concept of multi-level classification, which can significantly reduce the level of confusion. However, this model is prone to overfitting which needs careful pruning or setting depth of tree are essential to avoid this. The parameter that has been considered while constructing the decision tree is max_depth which denotes the maximum depth of the tree and is set to 6. Overfitting is caused by a larger maximum depth value, whereas underfitting is caused by a lower maximum depth value.

3.1.4 Random Forest (RF)

Random Forest is a supervised learning algorithm that selects the best prediction from numerous individual decision trees using ensemble learning. So, this multitude of uncorrelated decision trees work individually on various sub-samples of the dataset to perform a prediction and the best of which can outperform any individual constituent model through better predictive performance. These individual predictions are averaged to improve prediction and control over-fitting. The Random Forest classifier is mainly used for large datasets, and since RAVDESS dataset is also quite large, this classifier has been used. The parameters that have been considered for the random forest classifier are: n_estimators (default value = 100), random_state (value = 0).

A decision tree is generated on a whole dataset, using all of the relevant features/variables, whereas a random forest selects observations/rows and certain features/variables at random to build numerous decision trees from, then averages the results. So, both the classifiers have been used in this study to provide a better comparison.

The confusion matrices for the Decision Tree, Random Forest, Support Vector Machine, and Multi-Layer Perceptron techniques show that the anger emotion had the highest precision, implying that it was convenient to recognise. Neutral feelings, on the other hand, was the most challenging for all of the schemes to detect. Unlike our other classifiers, the MLP method relies on an artificial neural network (ANN) for classifying, and it delivered the highest precision in recognising emotions when likened to others [18].

3.2 Dataset

The most important element in carrying out SER is selection of an appropriate dataset which provides reliable and valid emotional expressions. The focus is mainly on speech, so the need is to look for datasets having more variations in the said speech performed by multiple actors. In this paper, RAVDESS (Ryerson Audio-Visual Database of Emotional speech and song) has been used as the dataset which consists of 24 professional actors (12 male, 12 female) from Canada, in a neutral North American accent, vocalizing two lexically-related utterances. This dataset has been rated by 319 raters from the same region for testing the reliability and validation. The model has been evaluated for 4 different emotions: calm, happy, angry and disgust where calm and neutral are selected as threshold conditions. There are two levels of emotional intensity in each expression: normal and intense. Such intensity variations help in clear and unambiguous results due to facial and vocal expressions being able to be more accurately identified [12]. This allows emotion detection for identifying both intensive expressions and subtle changes. Each recording of an actor is available in three modality formats: Audio-only (16 bit, 48 kHz.wav), Audio-Video (720p H.264, AAC 48 kHz, .mp4), and Video-only (no sound). In this work, the audio recording with 60 wav files for each of the 24 actors for a total of 1440 files studied over all 4 classifiers have been used. The selection of this database is supported by the fact that it provides a large data set for training and testing making it particularly suitable for emotion classifiers using supervised learning algorithms such as all the ones used here including MLP, SVM, Random Forest and Decision Tree.

Some datasets, such as the Surrey Audio-Visual Expressed Emotion (SAVEE) Database, use performers that have had extensive training and are asked to demonstrate various human emotions; this type of dataset is regarded as an Actor-based dataset. Another type is the elicited dataset, which involves exposing a set of people to a range of emotional cases and recording their responses, such as the eNTERFACE' 05 Audio-Visual Emotion Database. The final type of dataset is spoken datasets, that are gathered from real-life situations in order to elicit genuine emotions, such as capturing callers to a customer call centre. In this work, it was required to find a suitable set of data that was readily available and contained the primary feelings; the dataset was also required to include a large number of actors instead of relying on one or two performers. As a result, the RAVDESS dataset was chosen. The RAVDESS dataset includes visual and audio data for 24 actors who were asked to sing and talk two lines ("Kids are talking by the door," "Dogs are sitting by the door") while displaying various human emotions [17].

3.3 Feature Selection and Extraction

Feature selection is the task of automatically or manually selecting the features that provide the maximum to the target variable or output that one is interested in. To reduce the computational cost of modelling and, in some cases, to improve the model's performance, the number of input variables should be reduced. Feature extraction is the process in which the input dataset is measured quantitatively to draw differences between different emotional states and classes. These features are extracted from the speech signals we have given as input. In this paper, 5 features have been extracted: MFCC, Mel Spectrogram, Chroma, Contrast and Tonnetz.

i) MFCC- The Mel-Frequency Cepstral Coefficients (MFCC) feature extraction technique is popular for extracting speech features, and current research strives to improve its performance. For security purposes, MFCC is used in speech recognition systems and so on. The Mel scale is a scale that is used to compare the perceived frequency of a tone to the frequency that can be measured. It modulates the frequency to more closely match what the human ear can hear.
ii) Mel Spectrogram- A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale. This belongs to the librosa package.
iii) Chroma- The chroma feature is a condensed descriptor that reflects a musical audio source's tonal component. This is part of the librosa package as well.
iv) Contrast- A current technique of proposing a contrast sensitive mechanism in auditory neural processing is to suggest an auditory contrast spectrum extraction algorithm, which is a relative representation of auditory temporal and frequency spectrum.
v) Tonnetz- This also belongs to the librosa package. Calculates the features of the tonal centroid (tonnetz) The perfect fifth, minor third, and major third are all represented as two-dimensional coordinates in this form, which uses the method of 1 to project chroma characteristics onto a 6-dimensional basis.

MFCC has been chosen as it is a simulation of the human hearing system that aims to artificially recreate the ear's operating principle, assuming the human ear is a reliable speaker recognition system. MFCC features found in the observed difference in the human ear's important bandwidths were used to keep the phonetically crucial aspects of the speech signal using frequency filters separated linearly at low frequencies and logarithmically at high frequencies [13].

3.4 Experimental Procedure

The RAVDESS dataset has been taken in this process. The necessary libraries such as the librosa (package which provides the necessary building blocks for music and audio analysis), soundfile, pickle (for serializing and de-serializing a Python object structure), NumPy (is a library that contains multidimensional array objects as well as a set of routines for processing them), and sklearn (includes a number of useful machine learning and statistical modelling techniques, such as classification, regression, and clustering) have been imported. After that, some of the emotions have been chosen for observation which are:'calm','happy','angry','disgust' out of the total eight available emotions in the dataset. Feature extraction has been executed to extract the features such as mfcc, chroma, mel and so on. The entire dataset has been divided into two parts: training set and testing set (20%).

Using the training and testing dataset, different classifiers have been applied in order to compare which model gives the best accuracy.

After this, the Telegram bot has been set up to integrate with the SER model. All the commands for this Telegram bot, named Ascertus, are displayed in the bot menu. The/start command is used to launch our application. It shows a brief description of our project. The /emo command retrieves a voice sample and displays its predicted emotion. Voice samples from telegram are typically saved as.ogg audio files. This is then fed into the MLP classifier (the classifier with the highest training accuracy) to determine the emotion. Finally, the /exit command is used to terminate the application.

The motive of the updater class of the telegram.ext package is to receive updates from Telegram and deliver those to the dispatcher. The dispatcher supports handlers for various types of data: Telegram updates, basic text commands, and even arbitrary types Handler classes such as CommandHandler and MessageHandler handle telegram commands and user input. The Python pickle module is required to use our classifiers in our Telegram bot application. Pickling is a method for converting a Python object (list, dictionary, etc.) into a character stream. The idea is that this character stream contains all of the information required to reconstruct the object in a subsequent Python script. Along with creating the pickle files, we begin writing the app.py script. It is the main Python file that will be used to run and manage the Telegraph bot's functionalities. It is preferable to set the Telegram API access token within the host environment rather than in the project's source code.

The bot menu displays all of the commands for this Telegram bot, Ascertus. Our application is launched using the /start command. As soon as the bot is launched, it displays a brief description of the project that has been developed in this paper. The /emo command retrieves a voice sample and displays the emotion predicted by the sample. Telegram voice samples are typically saved as OPUS audio files. An Ogg Opus

file is an audio file in the.ogg format, which is a lossy audio format designed for Internet streaming. This is then fed into the MLP classifier (the classifier with the highest training accuracy) to determine the emotion. Finally, the /exit command is used to terminate the application with a thank you message. A telegram.Message object has been used in each step to display the instructions to the user via the reply text function. Also, an error handler function is used in case there are any problems with the program flow.

A modest telegram chatbot has been developed here, using machine learning algorithms to assist the user in their emotion analysis. Some previous research articles have implemented speech emotion recognition models into chatbots, but one dedicated to a Telegram bot is uncommon. Because of Telegram's rapid growth, we are able to reach the widest possible audience for our initiative.

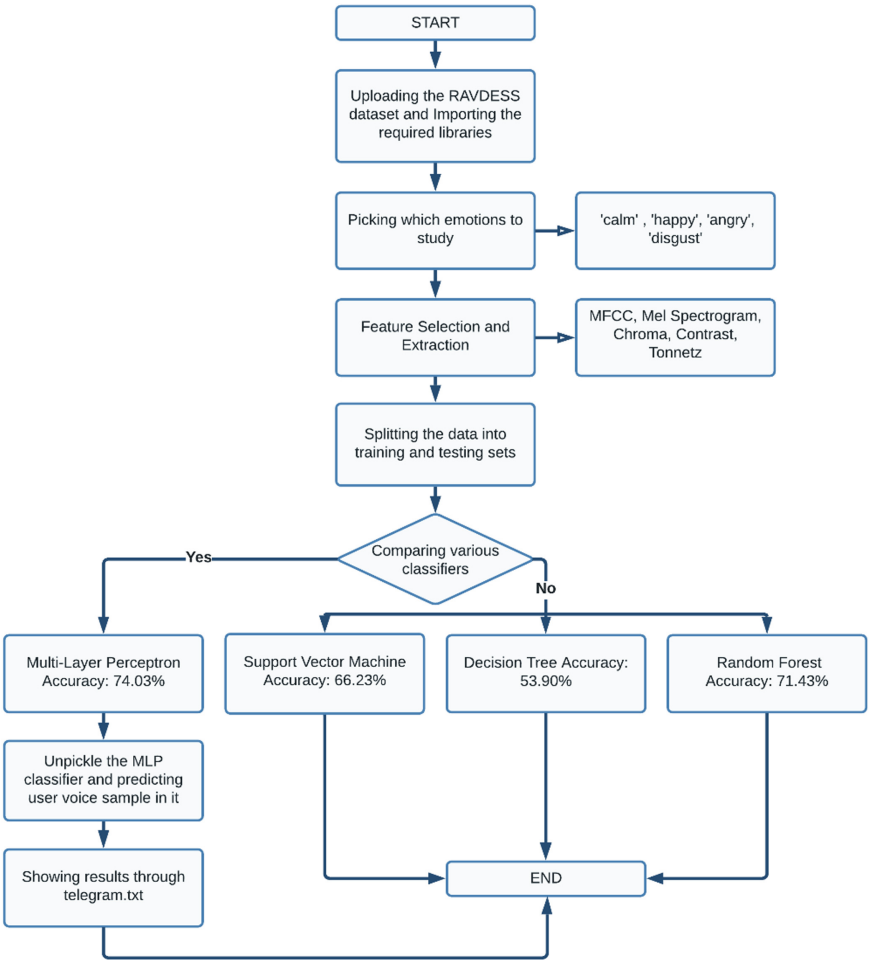

Fig. 1. Flowchart showing the workflow of the complete model.

The entire experimental procedure has been represented in a flowchart given below in Fig. 1.

4 Results

4.1 Accuracy

After carrying out the comparison of the different classifiers on the RAVDESS dataset for speech emotion recognition, a maximum accuracy of 74.03% has been achieved with the MLP classifier as shown in Fig. 2.

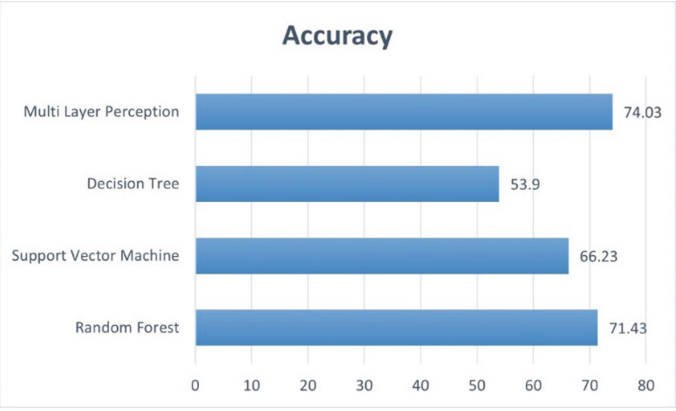

Fig. 2. Graph showing the accuracy obtained by various classifiers.

4.2 Classification Report

In machine learning, a classification report is a performance evaluation indicator. The precision, recall, F1, and support scores for the model are displayed in the classification report visualizer. The classification report for the various models that have been used in this project has been summarized in **Table 1.** as shown below. A graph has also been plotted for the same in **Fig. 3.**

4.3 Weighted Average

The weighted average shown in Fig. 4 takes into account how many of each class were used in the computation, having less of one class means that its precision/recall/F1 score has a smaller impact on the weighted average for each of those factors.

Table 1. Table showing the maximum value obtained for each observed emotion using all the four classifiers described in the paper.

	Maximum value of precision	Maximum value of recall	Maximum value of f1-score
Angry	Random Forest - 0.69	Multi-Layer Perceptron - 0.81	Random Forest - 0.70
Calm	Random Forest - 0.93	Multi-Layer Perceptron - 0.88	Random Forest - 0.90
Disgust	Multi-Layer Perceptron - 0.71	Random Forest, Support Vector Machine - 0.73,	Multi-Layer Perceptron - 0.64
Happy	Multi-Layer Perceptron - 0.81	Multi-Layer Perceptron - 0.64	Multi-Layer Perceptron - 0.71

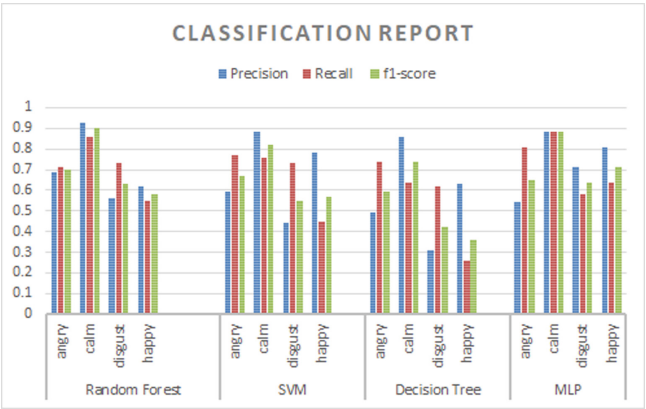

Fig. 3. Graph showing the comparison of various classification reports of all the classifiers used in the model.

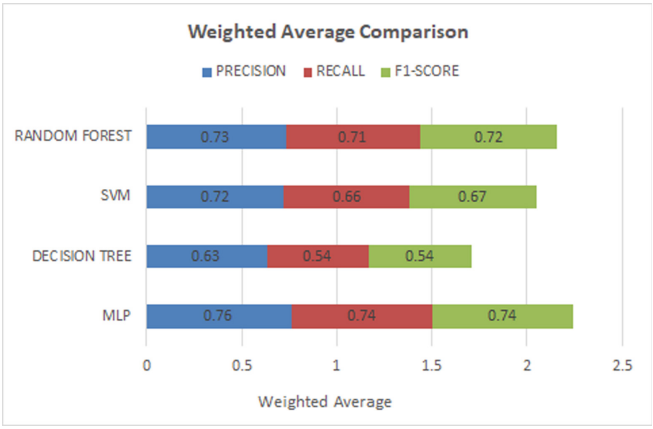

Fig. 4. Graph showing the average weighted comparison for the values of precision, recall and f1-score for the different classifiers.

4.4 Telegram Bot Integration

The integration of complex Machine Learning models into user-friendly applications like a chatbot requires effort in fine-tuning the model and adjusting it accordingly. The following image shown in Fig. 5 provides a glimpse of bot functionality performing basic emotion recognition. One can later build upon these produced results to draw inferences on more detailed data and perform diagnosis.

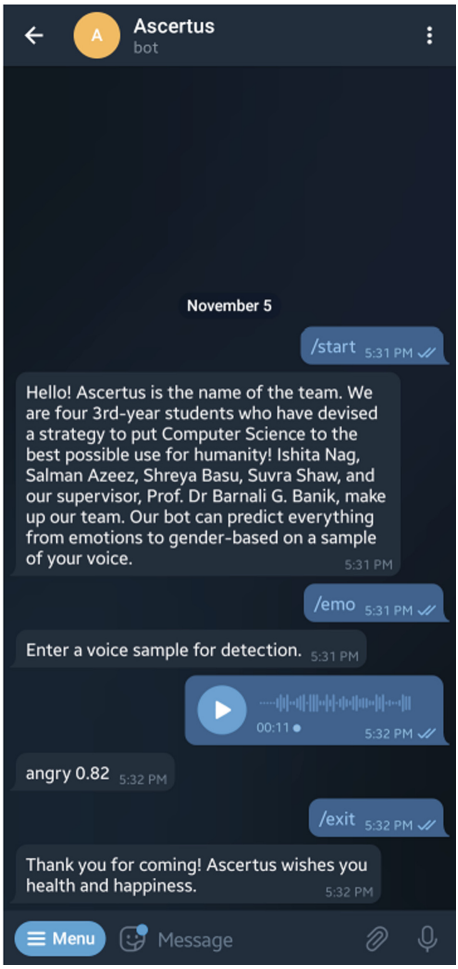

Fig. 5. Mockup of basic bot functionality

5 Discussions and Conclusions

Several methods have been researched in SER ranging from simple supervised models, deep learning models to even complex models with stacked classifiers. However, all these methods provide a specific and well-defined emotion detection algorithm by suitably modifying the parameter selection and fine-tuning the model for higher accuracies in particular scenarios. This paper aims for a more generalized model by providing comparative analysis of the most relevant yet simple classifiers over a wide range of selection features and incorporating it within a user-friendly chatbot. This telegram chatbot provides a more congenial environment which takes a step away from the enigmatic complex emotion recognition models. The idea for development of a chatbot comes from the fact that, unfortunately, mental conditions are not easily recognized or are frowned upon by the society. As a consequence, a mentally affected individual may not be aware of his condition, deducing it as stress or peer pressure. Even the aware may be conscience-stricken so as to not make it public. This mortification or false contrition can be justified as the world sees it so. Through this application of SER, the aim is to provide exposure to such situations, provide a user-friendly interface to realize, cope with their state of affairs and help improve the status quo. This accurate identification is also crucial for a person to be more self-aware, confident and have far better metacognition [14]. Although this paper highlights the foundations of this interactive psychotherapy python bot, work is still being conducted to include finer details as mentioned in the future works towards the end of this paper.

6 Future Work

This paper deals with simplistic yet reliable SER incorporated in a user-friendly telegram chatbot. However, when dealing with real-life situations, it calls for more robust algorithms that have close to none chances of providing inaccurate readings. Since only SER has been tackled, there is room for more multi-modal mechanisms through video and text inputs to deal with scored values. In addition to emotion, the next objective is to incorporate gender detection for even better fine-grained recognition and diagnosis. This allows for specialized and distinctive classification and therapy. As one can imagine, the application is far from complete and will require several iterations adding finer details in each step to reach somewhere close to completion. Hence, this paper marks the first step towards greater intelligence in psychological therapy.

References

1. Zhao, J., Mao, X., Chen, L.: Speech emotion recognition using DEEP 1d 2D CNN LSTM networks. Biomed. Signal Process. Control **47**, 312–323 (2019)
2. Milton, A., Sharmy Roy, S., Tamil Selvi, S.: SVM scheme for Speech emotion recognition USING MFCC feature. Int. J. Comput. Appl. **69**(9), 34–39 (2013)
3. Chenchah, F., Lachiri, Z.: Acoustic emotion recognition using linear and nonlinear cepstral coefficients. Int. J. Adv. Comput. Sci. Appl. **6**(11) (2015)
4. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Commun. **41**(4), 603–623 (2003)

5. Demircan, S., Kahramanlı, H.: Feature extraction from speech data for emotion recognition. J. Adv. Comput. Networks **2**(1), 28–30 (2014)
6. Le, D., Provost, E.M.: Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (2013)
7. Tawari, A., Trivedi, M.: Speech emotion analysis in noisy real-world environment. In: 2010 International Conference on Pattern Recognition
8. Shaqra, F.A., Duwairi, R., Al-Ayyoub, M.: Recognizing emotion from speech based on age and gender using hierarchical models. In: The 10th International Conference on Ambient Systems, Networks and Technologies (ANT), 29 April–2 May 2019, Leuven, Belgium (2019)
9. Nediyanchath, A., Paramasivam, P., Yenigalla, P.: Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
10. Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for Speech Emotion Recognition using convolutional Neural Networks. IEEE Trans. Multimedia **16**(8), 2203–2213 (2014)
11. Nalini, N.J., Palanivel, S., Balasubramanian, M.: Speech emotion recognition using residual phase and MFCC features. Int. J. Eng. Technol. **5**(6), 4515–4527 (2013)
12. Livingstone, S.R., 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), e0196391 (2018)
13. Alim, S.A., Alang Rashid, N.K.: Some Commonly Used Speech Feature Extraction Algorithms. Pub: 12th Dec, 2018
14. Lausen, A., Hammerschmidt, K.: Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. Humanities and Social Sciences Communications, vol. 7, no. 1 (2020)
15. Sanjita, B.R., Nipunika, A.: Speech Emotion Recognition using MLP Classifier. IJESC, vol. 10, no. 5, May 2020
16. Amjad, A., Khan, L.: Effect on speech emotion classification of a feature selection approach using a convolutional neural network. Peer J. Comput. Sci., **7**, Pub: 3rd Nov, 2021
17. Martin, O., Kotsia, I., Macq, B.: The eNTERFACE' 05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06); Pub: 24th Apr, 2006
18. Madhavi, A., Priya Valentina, A., Mounika, K., Rohit, B., Nagma, S.: Comparative analysis of different classifiers for speech emotion recognition. In: Kiran Mai, C., Kiranmayee, B.V., Favorskaya, M.N., Chandra Satapathy, S., Raju, K.S. (eds.) Proceedings of International Conference on Advances in Computer Engineering and Communication Systems. LAIS, vol. 20, pp. 523–538. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9293-5_48