Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021)
IEEE Xplore Part Number: CFP21N67-ART; ISBN: 978-0-7381-4627-0

Speech Emotion Recognition and Sentiment Analysis based Therapist Bot

Yashwardhan Bhangdia, Rashi Bhansali
Dept of Comp Engg, Vishwakarma Institute of Technology,
Savitribai Phule Pune University, Pune, India
yashwardhan.bhangdia18@vit.edu,
rashi.bhansali18@vit.edu

Ninad Chaudhari, Dimple Chandnani, M L Dhore
Dept of Comp Engg, Vishwakarma Institute of Technology,
Savitribai Phule Pune University, Pune, India
ninad.chaudhari18@vit.edu, dimple.chandnani18@vit.edu,
manikrao.dhore@vit.edu

2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) | 978-1-6654-3877-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIRCA51532.2021.9544671

Abstract* — Access to mental health care and treatment is a global concern. The demand for these services outnumbers the supply. Online treatment delivered by a chatbot might not only provide access to cost-effective aid, but it could also be convenient for individuals, who are hesitant to participate in therapy. The main aim is to leverage technology to enable more people to seek help. The proposed method has built a responsive therapist bot that generates appropriate responses according to a person's emotion. The detection phase uses an ensemble of deep learning models for Speech Emotion Recognition (SER) and text based sentiment analysis, which classifies one's emotions into four categories – happy, sad, angry, and anxious. The proposed approach can be proved to be fool-proof and it is better than existing methods due to the ensemble of two efficient models CNN (83%) and BiLSTM (92%). In addition to detection and responses, this application recommends suitable tasks to assist the target audience.** ***Keywords —Speech Emotion Recognition, Sentiment Analysis, Deep Learning, CNN, BiLSTM, Chatbot

I. INTRODUCTION

As per statistics from WHO, nearly 7.5 % of Indians suffer from different mental disorders. It is predicted that by the end of this year, close to 20 percent of India will suffer from mental illnesses. While anxiety and stress majorly affect people in the 15-29 age group, most of them do not seek professional help. This could be due to financial issues or mental health related stigma. The aim of this paper is to automate this process by creating an effective and fool-proof method to detect emotions. Therapist Chatbots have gained a lot of popularity over the recent years. However, existing systems in the market are text-based and use techniques like Cognitive Behavioral Therapy and meditation. While these have proven to be helpful, human emotions change almost every second and it is very challenging to articulate strong emotions into words. In fact, some emotions often go unnoticed in text-based systems.

To address this, a speech emotion recognition element has been added to capture the prosodic features and frequencies in one's voice that can improve the efficiency. Here, 4 major emotions have been considered - happy, sad, angry, and anxious as per Plutchik's Wheel of Emotions. This organizes 8 basic emotions based on the physiological purpose of each - anger, anticipation, joy, trust, fear, surprise, sadness, and disgust. The combination of Speech Emotion Recognition [SER] and text-based sentiment analysis gives the proposed approach with an edge over the others. This system also suggests relevant exercises for each emotion like deep breathing, listening to soothing music and others. The model has been deployed on a web-based application with a minimalist interface that follows UI principles.

II. RELATED WORK

The Plutchik's Wheel provides a logical approach for classifying emotions into eight sectors. Recognizing emotions through speech requires capturing the prosodic features like stress, tone, rhythm, and phonetics besides the acoustic features. Prosodic features are basically those features which appear in a complete sentence, or which reflect the meaning of the sentence through the form of utterance of the speaker i.e., lightweight, heavy weight or super heavy weight stress. As in [12], they show a comparative analysis of Mel Frequency Cepstral Coefficient (MFCC) as short term features and prosodic features as long term ones. However as in [1] they conclude that spectral features are dominant in predicting the emotions.

Emotions are fleeting, often overpowering, and difficult to interpret through a single way. As in [6], they show that speech expressions can be used for detecting emotions efficiently. Their approach uses Support Vector Machines (SVM) for classification. They have used the SAVEE dataset for audio-visual features and the RAVDESS dataset for speech features as given in [10]. They claim to have a model that gives 95.3% accurate results. As in [2], they used SVM as a classifier to identify different emotions mainly rage, sadness, fear, happiness and neutral. They used the Berlin database of emotions and LIBSVM for classifying emotions. They achieved an accuracy of 94.7% for male audio and 100% for female speech while an overall accuracy of 93.7% was obtained for gender independent cases.

As in [13], they used a unique Convolutional Neural Network (CNN) based speech-emotion recognition system. They processed every audio file with 1D CNN layers. Their system had 83.61% accuracy. As in [14], they tested various Machine Learning paradigms followed by Recurrent Neural Networks, which they used to classify seven emotions. An accuracy of 94% was obtained with the Spanish database, using the RNN classifier with feature selection. As in [15], they focus on an attention mechanism. Their approach of

relation aware attention-based 3D CNN and LSTM model achieved an accuracy of 80.80%.

TABLE I.I RELATED WORK - SPEECH EMOTION RECOGNITION

Reference Paper	Features	Method	Result
Class-Level Spectral Features for Emotion Recognition[1] (2010)	MFCC (spectral features) & phoneme type classes -prosodic features	Phoneme level segmentation, classification	Prosodic 68.1% Spectral 72.5%
SER using SVM[2] (2010)	MFCC and MEDC	SVM for emotion detection	93.7%
Automated SER on Smartphones[6] (2018)	MFCC	SVM	95.3%
CNN based SER [13] (2019)	MFCC	CNN	83.6%
Automatic SER using Machine Learning[14] (2019)	Modulation Spectral Features, MFCC	RNN	94%

Emotion detection is a part of sentiment analysis as words are the most effective medium of expressing emotions. As in [4], they achieve excellent results on sentence classification through hyperparameter tuning and static vectors. As in [7], they demonstrated that a feed-forward network with a simple globally constrained decoder can accurately and quickly annotate several languages pairs of code-mixed and single-language texts for vernacular languages. They conclude to have achieved a 19.5% averaged absolute gain. As in [9], they performed text classification based on LSTMs and attention mechanisms to improve the text sentiment classification. As in [16], they showed that Bidirectional LSTMs perform better in sentiment analysis as they analyse words based on preceding and succeeding words. They used the IMDB dataset and their model achieved an accuracy of 88.5%.

TABLE I.II RELATED WORK – SENTIMENT ANALYSIS

Reference Paper	Features	Method	Result
Convolutional neural networks for sentence classification[4] (2014)	Static pre trained word vectors	Simple CNN, tuning for sentence level classification	81.5%
A fast, compact, accurate model for language identification of codemixed text[7] (2018)	Character, script, and lexicon features	SVM for emotion detection	92%
Text classification based on LSTM and attention[9] (2018)	Local features of the text extracted from CNN	CNN for feature extraction, LSTM and Attention based mechanism for classification	86%
Aspect Level Sentiment Analysis Bi-Directional LSTM Encoder with the Attention Mechanism [16] (2020)	GloVe embeddings for pre trained vectors	BiLSTM with attention mechanism for polarity of sentences	88.5%

III. METHODOLOGY

The overview of the architecture of the proposed system is depicted in figure 1.1. It mainly involves following components:

a. Speech Emotion Recognition that involves extraction of the spectral and prosodic features from the input audio signal.
b. Speech to text conversion of the audio signal followed by structuring, cleaning, and preprocessing the text data.
c. Ensemble of Deep Learning Models for emotion classification based on Mel-frequency features extracted in step a, and sentiment analysis of text generated in step b.
d. Development of an interactive chatbot application that generates appropriate responses based on analysis done on user input.

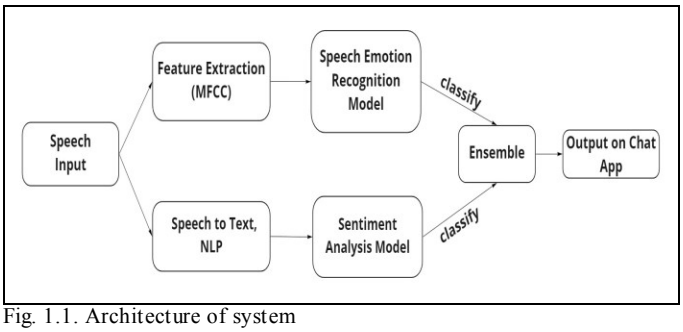

Fig. 1.1. Architecture of system

The algorithm in figure 1.2. demonstrates the flow of the chatbot system implemented in the proposed research.

A. Speech Emotion Recognition

Dataset

For the purpose of detecting emotions from speech, labelled data with about 7326 audio files has been collected from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset found in [10], a labelled data set that has audio files in waveform format for eight emotions - neutral, calm, happy, sad, angry, fearful, disgust, and surprise. Out of these, 786 files have been chosen for the selected list of emotions, namely, joy, sadness, anger, and fear. About 75% of data is used for training and 25% is used for testing the SER models.

Feature Extraction - MFCCs

The data is in waveform format, which has a sampling frequency of 44,100 Hertz. In order to extract quantitative features from audio signals, the librosa library in Python is used. It gives 40 Mel-frequency Cepstral Coefficients (MFCCs) representing energy levels in an audio essential for capturing the important frequencies. The process of extracting MFCCs involves the following steps as elaborated in [5]:

a. Resampling and Frame Segmentation

The audio signal which represents the variation in air pressure over time, is resampled at a frequency of 22050

Hertz, that is half of the original frequency. This frequency is decided according to the Shannon-Nyquist Theorem as stated in [18]. In order to process the signal, it is segmented at short time intervals corresponding to a default value of 2048 overlapping samples. This gives a frame length of 93 milliseconds as given by equation (1). Hop length is 512 milliseconds.

$$T=N/Fs \qquad \text{eq(1)}$$

where T is the frame length in seconds, F_s is the sampling frequency in Hertz and N is the number of samples.

b. Fast Fourier Transform

Discrete Fourier Transform is a mathematical formula represented by equation (2), which decomposes a signal varying in time, into its composite frequencies. It maps each frame of N samples of input signal s(n), from time domain to frequency domain thereby generating a spectrum. Fast Fourier Transform is a fast and efficient method for generating spectrograms computed on overlapping segments of signal.

$$S_k = \sum_{n=0}^{N-1} s(n) e^{\frac{-2i\pi nk}{N-1}} \qquad \text{eq(2)}$$

a. Mel frequency

The following equation (3) is used to convert the spectrum obtained from FFT to Mel-scale. This step is performed because the human ear can differentiate between frequencies in the lower range better than those having the same difference in the upper range.

$$M(f) = 2595 log_{10}(1 + \frac{f}{700}) \qquad \text{eq(3)}$$

The Mel Filter bank consists of 40 triangular filters which are applied to the spectral estimate. Finally, Discrete Cosine Transform is applied to the 40 logarithmic filter bank energies which give the 40 coefficients for emotion recognition.

Evaluating Emotion in Speech Signal

Out of the 40 MFCCs extracted by the process mentioned above, only lower 13 are useful for identifying the emotions. For mapping these values to the corresponding emotion classes, following ranges of standard deviation of MFCCs, for each emotion are obtained experimentally, as shown in table II.

TABLE II. MFCC TO EMOTION MAPPING

Sr. No.	Emotion	MFCC - Standard Deviation Range
1	Joy	2.81 - 2.93
2.	Sadness	3.01 - 3.12
3.	Anger	2.73 - 2.79
4.	Fear	2.037 -2.51

Algorithm 1 Algorithm for Proposed System

Input: Speech (audio sampled at 44.1kHz in wave format)
Output: Chat Response based on Emotion detected
initialization : endChat = False
LOOP Process
while *endChat* $\neq$ *True* **do**
  Compute $MFCC[]$ = MFCCs from speech input
  Compute $text$ = speech input to text
  Compute $P_s[]$ = Classification Probabilities by CNN SER Model($MFCC[]$)
  Compute $P_t[]$ = Classification Probabilities by BiLSTM Sentiment Model($text$)
  Compute $maxP_s = \max(P_s)$, C_s = emotionClass($maxP_s$)
  Compute $maxP_t = \max(P_t)$, C_t = emotionClass($maxP_t$)
  if $maxP_s$ ¿ $maxP_t$ **then**
    $emotion = C_s$
  end if
else
  $emotion = C_t$
  if $emotion$ = happy **then**
    $response$ = bye
    $endChat$ = True
  end if
else
  $response$ = Generate response based on emotion from pre-determined tree of responses
  print $response$
  accept speech input
end while
return $response$

Fig. 2.2. Algorithm of Proposed System

Machine Learning and Deep Learning

About 40 features extracted from the training set were fed to a couple of machine learning algorithms like Decision Trees and Support Vector Machines (SVM). These algorithms are known to perform well on high dimensional data. However, the machine learning models did not converge well on the data. A maximum of about 58% accuracy could be achieved. To overcome underfitting, several deep learning models were used, namely, Multi-level Perceptron and 1-D Convolutional Neural Networks. About 83% accurate results were achieved with deep learning as shown in table III.

TABLE III. COMPARATIVE STUDY OF ALL ALGORITHMS

Sr. No.	Algorithm	Classification Accuracy (%)
1	Decision Trees Classifier	55.1%
2.	Support Vector Machines	57.8%
3.	Multi-Level Perceptron	61.9%
4.	Convolutional Neural Networks	83%

Classification and Regression Trees (CART), a decision tree classifier, was used to classify the data based on Gini index split for attributes. The minimum number of samples in every leaf node was tuned to a value of 20 to produce the best accuracy of 55% on the given data. After this, the SVM model was trained using a linear kernel to give an accuracy of 57.8% similar to the approach used in [2]. Deep learning architecture of Multi-Level Perceptron with 1 hidden layer made up of 300 neurons, a batch size of 256 produced 62% accurate results.

Convolutional Neural Networks

Convolutional Neural Networks, conventionally used on image data, were used to classify audio features. This choice was inspired by the one used in [13], where every audio data will have the features in a single dimension as opposed to images where there are three or four channels. Hence, 1D CNN layers were sufficient to take the audio vectors as input, assign weights and biases to them that would be learned by backpropagation. After a series of kernel based operations, max pooling, and activation, and finally feeding them to a feed forward vanilla neural network, the hyper parameters are tuned according to selected data. It is a multilabel classification that involved predicting the probabilities of all four types of emotions, the output layer consisted of 4 neurons with a sigmoid activation function for each instead of one neuron and softmax activation. With meticulous analysis, the number of hidden layers was chosen to be 3 and appropriate learning rate, epochs and sparse categorical cross entropy were chosen for loss function minimization and about 83% accuracy was achieved and good f1-score as shown in the confusion matrix in figure 2 where 0, 1, 2, and 3 are class labels for "joy", "sadness", "anger" and "fear" respectively. As a result, CNNs were used in the final deployment.

	precision	recall	f1-score	support
0	0.82	0.69	0.75	52
1	0.90	0.90	0.90	43
2	0.73	0.89	0.80	61
3	0.89	0.89	0.89	57
accuracy			0.83	213
macro avg	0.83	0.84	0.83	213
weighted avg	0.83	0.83	0.83	213

Fig. 2. Confusion Matrix for SER using CNN Classifier

B. Sentiment Analysis based on Text

Dataset

This phase of the research involved detecting emotions based on text. A labelled dataset was scraped from twitter consisting of about 20,000 tweets extracted using the Twitter API. It included six labels, namely, Anger, Joy, Sadness, Fear, Surprise and Love. These six labels were then mapped to the four emotions mentioned in the speech emotion recognition section while taking the ensemble. About 16000 records were used for training and 2000 for testing and validation each, that is a 75-25 train-test-validation ratio as shown in figure 3.

Preprocessing

As the dataset consisted of a variety of sentences including slang words which were not useful in analyzing the sentiment of a person. A lot of punctuation marks and other special symbols were present in the data which were supposed to be removed for the data to be clean before it is fed to the model. The data underwent many preprocessing steps where the first step was to remove all the redundant punctuation marks and stop words in English such as "the", "is", "this", etc. The python text processing library NLTK was used for removing the stop words while for removing the punctuations simple regular expressions were used. The sentences in the dataset were then tokenized followed by lemmatizing individual words as a part of normalizing the data where words like "running" are brought to their root word "run" ensuring that the root word belongs to the language.

The deep neural models needed text embeddings. Embeddings basically capture the representation of the word in a higher dimensional plane through which a vector representation of the words was created. The word2Vec library from the *gensim* package was used to achieve this by training it on our own text corpus.

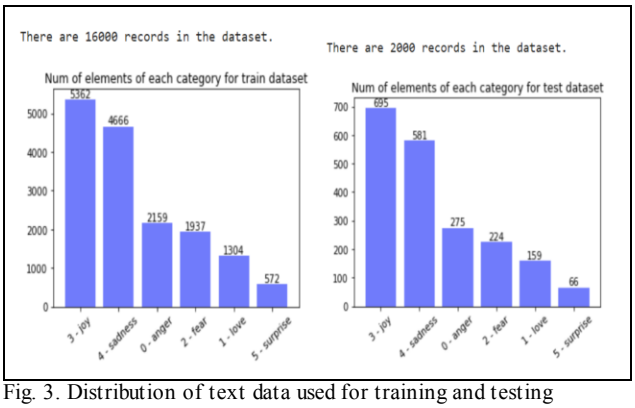

Fig. 3. Distribution of text data used for training and testing

Deep Learning (Bi-directional Long Short Term Memory)

Once the textual data was preprocessed it was fed to a deep learning model namely Bidirectional Long Short-Term Memory (BiLSTM) network. In the LSTM network [9], the input is read only in forward direction but to capture the semantic meaning of a sentence the subsequent words are of equal importance. A BiLSTM architecture similar to the one in [11] was used to improve the contextual understanding of the neural network. It was a sequential model consisting of an Embedding layer, Bidirectional LSTM Layers, Dropout layers and the final dense feed forward neural network.

The embedding layer was fed with custom embeddings created during preprocessing of the text. The hidden layers were chosen to be 3 LSTM networks after careful inspection and the default hyperbolic tangent (tan h) function was used in all the layers. Subsequently, after every hidden layer a dropout layer was added to prevent the overfitting of the model. Since it was a classification for 6 different types of sentiments, the output layer consisted of one neuron with a softmax function. After choosing the appropriate number of epochs and optimization functions a maximum of 92% accuracy was obtained as shown in the confusion matrix in figure. 4. where 0, 1, 2, 3, 4, and 5 are class labels for "joy", "sadness", "anger", "fear", "surprise" and "love" respectively.

C. Ensemble of Classifiers

Ensemble learning is an optimization technique used to improve model accuracy which involves building a new classifier from existing heterogeneous classifiers i.e., the base classifiers can use different algorithms, tuning parameters, and

training sets. The goal is to identify the best fitting parameters of all the base classifiers and build a new one with reduced bias and variance. For the purpose of this research, individual classifiers (CNN and BiLSTM) were combined using the method of majority voting called as soft voting that involved averaging predicted probabilities for different class labels and predicting the one with the largest average. Detection of emotions from words alone can often give the wrong idea. The pitch, loudness, and acoustic as well prosodic features of one's voice also say a lot about one's emotions. Thus, an ensemble of two techniques was generated for more precision.

	precision	recall	f1-score	support
0	0.96	0.92	0.94	695
1	0.92	0.94	0.93	275
2	0.75	0.85	0.80	159
3	0.96	0.97	0.97	581
4	0.94	0.82	0.88	224
5	0.67	0.89	0.77	66
accuracy			0.92	2000
macro avg	0.87	0.90	0.88	2000
weighted avg	0.93	0.92	0.92	2000

Fig. 4. Confusion Matrix for Sentiment Analysis using BiLSTM classifier

D. Development of the Chatbot Web Application

Graphical User Interface was of utmost importance to us in terms of keeping the user engaged throughout their interaction with the chatbot. HTML was used for creating a basic layout of the web application along with CSS and Bootstrap for styling. The primary color used was green which is known to have a balancing and harmonizing effect on users helping them diffuse anxiety, stay calm and refreshed. JavaScript was used to send audio files to the backend by making real time post requests to the server route. Figure 5 shows the chat application layout.

Fig. 5. Graphical User Interface of Chat Application

To deploy the ensemble, the model was integrated with the backend of the web application. A HDF5 file containing the learned weights and parameters of the trained model, was used for real time prediction of emotions based on the speech input provided by the user. This was done using Flask which is a lightweight web application framework which does not require particular tools or libraries. Once the emotion was predicted, a suitable response was generated based on a response tree that drew inspiration from the Cognitive Behavioral Therapy. CBT focuses on identifying and challenging automatic negative thoughts which often make a person depressed or anxious and replaces these thoughts with more pragmatic ones. Suitable activities were recommended to provide instant distraction from such thoughts as shown in figure 5.

An end-to-end testing was performed for checking the compatibility of all the components used in the web application and also for integrating the front-end and the back-end.

IV. RESULT ANALYSIS

Proposed methodology was tested and evaluated against different models. The accuracy of CNN for speech emotion recognition was calculated to be 83% against the RAVDESS dataset. The accuracy of BiLSTM Model for sentiment analysis from text was calculated to be 92% against the Twitter corpus. Since both models are trained on different data, the ensemble is heterogeneous. However, while deploying, probability of classification is calculated by both methods for the same speech input, converted to text for the latter model. Maximum of the two probabilities is taken to ensure accurate results.

Table IV depicts the comparative experimental study of proposed model with related work that has been done previously for speech emotion recognition in [3], [13] and [15]. The proposed model performs better than the existing systems.

TABLE IV. EXPERIMENTAL RESULTS COMPARISON FOR SPEECH EMOTION RECOGNITION

Reference Paper	Methodology	Accuracy
Speech Emotion Recognition[3]	SVM	81%
Convolutional Neural Network (CNN) Based Speech-Emotion Recognition[13]	MVR SVM RNN CNN	57% 66% 68% 83%
Speech Emotion Recognition Using Convolutional Neural Network and Long-Short Term Memory [15]	3D CNN LSTM	82%
Proposed paper results Speech Emotion Recognition	CNN	83%

Table V shows the comparison of proposed method with the approaches used in [8], [16] and [17] for sentiment analysis. It also calls attention to the value addition given by proposed method with 83% and 92% accuracy scores, respectively.

Although, several machine learning and deep learning models with good performance have been trained and

deployed before, it was found that the proposed model gives better results in terms of both precision and accuracy. Besides, the proposed method is instrumental in building a practical and potentially nifty chat application for therapy automated with the help of an ensemble of the exceptionally performing models.

TABLE V. EXPERIMENTAL RESULTS COMPARISON FOR SENTIMENT ANALYSIS

Reference Paper	Methodology	Accuracy
A topic BiLSTM model for sentiment classification[8]	Topic Information-Based Bi-LSTM	95%
Aspect Level Sentiment Analysis Using Bi-directional LSTM Encoder with the Attention Mechanism[16]	Bi-directional LSTM	88.5%
Sentiment Classification Using a Single-Layered BiLSTM Model[17]	Single layered Bi-LSTM	85.78%
Proposed paper results Sentiment Analysis	Bi-directional LSTM	92%

V. CONCLUSION AND FUTURE SCOPE

A Multi-modal emotion detection based therapy chatbot was developed and successfully deployed on a web application. The chatbot successfully detects correct emotions, maintains appropriate conversation replies accordingly and also suggests some engaging activities for particular emotions based on a person's mood. The superiority of this project lies in the fact that it uses an ensemble of audio based SER and text based (Sentiment analysis) models, which results in the chatbot predicting emotions with better accuracy than the traditional text based chatbots. The final outcome of this research was a foolproof conversational chatbot with highly accurate results.

An additional feature to this project can be the inclusion of facial recognition as well, for detecting emotions. This might produce better results as an ensemble of three different factors will be considered. The responses generated by the chatbot can be automated instead of a conversation tree by using recently developed models like GPT-3 or RNNs where the user's conversation can be understood by the chatbot semantically. The neural network backend slows down the response time of the chatbot, which can be reduced by making a remote request call to a hosted API. One major advancement that can be made in this project is to increase the conversation scope of the chatbot to various languages.

REFERENCES

[1] Bitouk, Dmitri et al. "Class-Level Spectral Features for Emotion Recognition." Speech communication vol. 52,7-8 (2010): 613-625. doi:10.1016/j.specom.2010.02.010

[2] Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Computer Applications, vol.1, pp.6-9,February 2010..

[3] S. Lalitha, A. Madhavan, B. Bhushan and S. Saketh, "Speech emotion recognition," 2014 International Conference on Advances in Electronics Computers and Communications, 2014, pp. 1-4, doi: 10.1109/ICAECC.2014.7002390

[4] Y. Kim, ''Convolutional neural networks for sentence classification,'' in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 1746–1751.

[5] Likitha, M. S.; Gupta, Sri Raksha R.; Hasitha, K.; Raju, A. Upendra (2017). [IEEE 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) - Chennai, India (2017.3.22-2017.3.24)] 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) - Speech based human emotion recognition using MFCC. , (), 2257–2260. doi:10.1109/WiSPNET.2017.8300161

[6] H. Alshamsi, V. Kepuska, H. Alshamsi and H. Meng, "Automated Speech Emotion Recognition on SmartPhones," 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2018, pp. 44-50, doi: 10.1109/UEMCON.2018.8796594

[7] Y. Zhang, J. Riesa, D. Gillick, A. Bakalov, J. Baldridge, D. Weiss, "A fast, compact, accurate model for language identification of codemixed text", 2018, pp. 328–337. doi:10 .18653/v1/D18-1030

[8] Yanming Huang, Y. Jiang,Touhidul Hasan,Q. Jiang,Chao Li ,"A topic BiLSTM model for sentiment classification," in ICIAI '18: Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence, 2018,https://doi.org/10.1145/3194206.3194240.

[9] X. Bai, "Text classification based on LSTM and attention," 2018 Thirteenth International Conference on Digital Information Management (ICDIM), 2018, pp. 29-32, doi: 10.1109/ICDIM.2018.8847061.

[10] Livingstone SR, Russo FA (2018) 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. https://doi.org/10.1371/journal.pone.0196391.

[11] G. Xu, Y. Meng, X. Qiu, Z. Yu and X. Wu, "Sentiment Analysis of Comment Texts Based on BiLSTM," in IEEE Access, vol. 7, pp. 51522-51532, 2019, doi: 10.1109/ACCESS.2019.2909919.

[12] A. B. Gumelar et al., "Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks," 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), 2019, pp. 1-8, doi: 10.1109/SeGAH.2019.8882461.

[13] A. B. Abdul Qayyum, A. Arefeen and C. Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), 2019, pp. 122-125, doi: 10.1109/SPICSCON48833.2019.9065172.

[14] Kerkeni, Leila and Serrestou, Youssef and Raoof, Kosai and Cleder, Catherine and Mahjoub, Mohamed and Mbarki, Mohamed, "Automatic Speech Emotion Recognition Using Machine Learning.", March, 2019.

[15] Dangol, R., Alsadoon, A., Prasad, P.W.C. et al. Speech Emotion Recognition Using Convolutional Neural Network and Long-Short Term Memory. Multimed Tools Appl 79, 32917–32934 (2020). https://doi.org/10.1007/s11042-020-09693-w

[16] Kay Khine W.L., Thwet Aung N.T. (2020) Aspect Level Sentiment Analysis Using Bi-Directional LSTM Encoder with the Attention Mechanism. In: Nguyen N.T., Hoang B.H., Huynh C.P., Hwang D., Trawiński B., Vossen G. (eds) Computational Collective Intelligence. ICCCI 2020. Lecture Notes in Computer Science, vol 12496. Springer, Cham. https://doi.org/10.1007/978-3-030-63007-2_22

[17] Z. Hameed and B. Garcia-Zapirain, "Sentiment Classification Using a Single-Layered BiLSTM Model," in IEEE Access, vol. 8, pp. 73992-74001, 2020, doi: 10.1109/ACCESS.2020.2988550.

[18] https://en.wikipedia.org/wiki/Nyquist_frequency