Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 4, 10479–10485

Deep Learning Methods for Suicide Prediction using Audio Classification

Dr. P. Golda Jeyasheeli[1], C. Kamaleshwar[2], K.Sakthi Aswin[2]

Department of Computer Science and Engineering
Mepco Schlenk Engineering College, Sivakasi
[1]-Professor (pgolda@mepcoeng.ac.in),
[2]-B.E Final year Students (kamaleshkamalesh955@gmail.com,aswin.offl08@gmail.com)

Abstract

Screening the suicidal ideation of people is one of the highly essential needs in this fast-moving depressing world. We aim to design a model for finding suicidal ideation based on the context spoken in an audio file. The proposed models are trained using RAVDESS and TESS audio emotion datasets. Features in the audio files are obtained using MFCC extraction. The various models trained and compared are: Random Forest (RF), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Dense Neural Network (DNN) on the extracted MFCCs. The audio files are classified into positive, negative and neutral emotion audio. DNN model produced 87% accuracy. The neutral and negative emotion audio files are further processed with the audio tokenizer model Wav2Vec to generate text transcriptions. The generated text transcriptions are classified with RoBERTa Model, trained on suicide depression dataset from Kaggle, which will classify the given audio into Suicide or Non-Suicide. The RoBERTa Model achieves 98% accuracy in classification.

Keywords: LSTM, RoBERTa, Deep Learning, Transformers, Speech Emotion Recognition (SER), Suicide Ideation Detection(SID)

1. INTRODUCTION

Suicide prevention is a challenging issue in a modern world. According to the statistics, for every hour, a student commits suicide in India. National Crime Records Bureau reports said that almost 28 suicides are being committed every day. The suicide rate is increasing gradually year after year. The global report shows that 8 lakh people commit suicide every year. The suicide rate among male is 3.4 percent higher than the female. Screening the risk among people for suicidal ideation and early detection of suicide can by most probably prevent suicide. Internet contains content in text, audio, images and video etc. All the different types of content may express suicide intent by any way like a suicide note etc. We have developed deep learning models which could detect the suicidal intent in the content posted online. In this paper, the audio content is analyzed for suicidal intent and then classified into either suicidal or non-suicidal intent. Training on data from users also invoke ethical issues like privacy which can be done by getting consent over the platform. Also, other ethical issues like bias while training on a particular set of data must also be dealt properly. The audio considered is processed to generate transcripts which could be analyzed for suicidal intent detection. For intent detection in text, transformer models BERT, ELECTRA and RoBERTa are trained on Kaggle dataset Suicide Depression detection. The algorithms used are evaluated on F1 score and accuracy. The application of the work is to create an API which could be used for automatic screening of social media posts by people with suicidal intent in real time.

2. LITERATURE REVIEW

Pavankumar Dubagunta et. Al. [1] worked on a deep Learning approach like Convolution neural Network (CNN). They give a voice input sample to detect depression using features from Raw Speech like Low Pass Filter (LPF), Linear Prediction Residual (LPR), Homomorphically filtered voice source and Zero frequency Filtered. Then these features are passed to the CNN Model and they used MFCC extraction to eliminate the Vocal Tract System. They find depression score and with it, depression state is predicted.

Anas Belouali et. al. [2] worked on acoustic and linguistic features together. They worked on a collected dataset from military voice calls and annotated suicide and non-suicide intent. For acoustic features, prosodic, phonetical and glottal features such as MFCC, energy entropy, zero crossing rate etc. are considered as features for the model. For linguistic features, LIWC, sentiment analysis was done along with parts of speech tagging. These features were considered either as standalone as acoustic features only, linguistic features only and with using a combined feature set of acoustic and linguistic features. The models used for comparison are Random Forest, SVM, Logistic Regression, XGBoost and DNN.

Stefen Scherer et. al. [3] worked on Machine Learning Methods like Support Vector Machine and Hidden Markov Model (HMM). By using HMM, they classify each frame in a 100Hz feature vector. The SVM is trained on standard deviation and median. They con- ducted two analysis of speech such as Interview level analysis and utterance level Analysis. On Comparison, Hidden Markov Model (HMM) gives a better performance than SVM in both interview and segment analysis.

Md Nasir et. al. [4] worked on Multimodal Interaction Model in a couple conversation. They used a Support Vector machine for classification of depression stages. They detected the speech using Diarization and Automatic Speech Recognition. After that, they extracted the features using methods such as Acoustic low-level Descriptors, Acoustic Behavior Embeddings, Lexical Cues, and Turn Taking Cues. Finally, classification using the Support Vector Machine was done.

Bhanusree Yalamanchili et. al. [5] worked on Machine Learning Methods like Support Vector Machine (SVM), Logistic Regression, and Random Forest. They worked on the features extracted using a COVAREP toolbox and each frame considered was sampled in 10ms. They applied PCA (Principal Component Analysis) to reduce the dimensions of features to eliminate the curse of dimensionality. For classification they used a SMOTE analysis to improve the accuracy. All the machine learning models are implemented in the scikit-learn toolbox. After extracting the feature and preprocessing the data was classified. SVM classifier is used to classify after performing SMOTE on the data to reduce the dimensional complexity.

Likhita M. et. al. [6] performed audio classification based on emotion. The researchers obtained MFCC after windowing and Fast Fourier transforms. After that the MFCC is created by converting frequency to Mel Scale. Average of the MFCC is taken and the standard deviation is measured which is passed to the classifier model which classifies based on the emotion.

3. PROPOSED WORK:

Here we discuss the process of the proposed work and how the data will flow in each module. Figure 1 represents the system design of the proposed work.

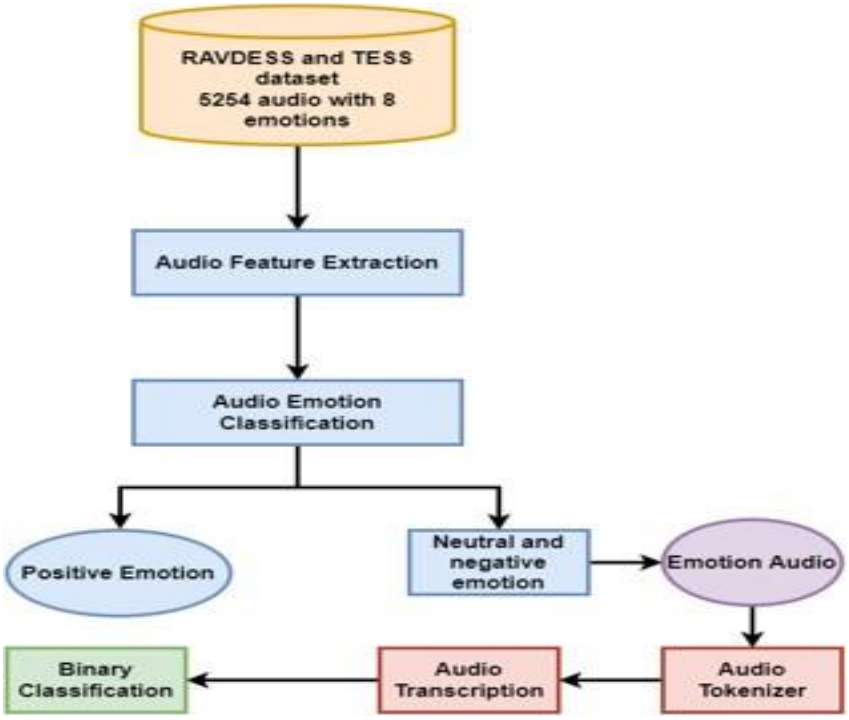

Figure 1: **System Diagram of The Proposed Model**

3.1. Dataset Description

For audio classification, RAVDESS & TESS dataset are used. They contain a total of 5254 audio files whose emotion is annotated as labels which are used for training and testing the model for classification. Each audio file of the dataset expresses an emotion like anger, sadness, fear, etc. which were manually annotated. For text classification, the Suicide and Depression Detection dataset, from Kaggle is used. The dataset has two columns: "text" which represents the posts' content, and "class" which represents the posts' label. This dataset is a collection of 232,074 posts from Reddit with equal suicide and non-suicide posts.

3.2. Audio Feature Extraction:

Every speech made by humans contains glottal pulses and vocal tract frequencies combined. First, the audio is converted from analog to digital by sampling. With the help of high pass filters, the process of boosting frequencies to enhance the phonetic features detection accuracy, known as pre-emphasis is done. Then the audio is divided into frames of 25ms for processing later. Then Windowing and Discrete Fourier transformations are applied to convert the time domain to the frequency domain. The frequency from each window or frame is converted to mel scale by applying the formula.

mel= 1127.01048 * log(f/700 +1) ,where f is the frequency to be converted

Then by applying Inverse Discrete Fourier Transform (IDFT), vocal tract frequency responses can be separated from glottal pulses. Then the Direct Cosine Transform is applied to the resultant cepstrum to decorrelate the values as obtained. 40 MFC coefficients are extracted from the audio which is used for the emotion classification. MFCC extraction is done with the help of the Librosa module in python.

3.3 Audio Emotion Classification

With the help of the features extracted, the following were trained: Random Forest, Dense Neural Network, Long Short-Term Memory, and Convolutional Neural Network models to classify the audio into Happy, Surprised, neutral, calm, angry, disgust, fear, sad. From this, we consider the neutral and negative emotions i.e., emotions except happy and surprised for identifying the suicidal indent in speech.

3.3.1 Random Forest

Random Forest is a Tree based ensemble learning method. Scikit learn's Random Forest Classifier was used with default hyper parameters. The model was trained with 80% of the data and tested on the rest. The model's performance was estimated on the standard metric terms accuracy and f1 score.

3.3.2 CNN

Convolutional Neural Network (CNN) was used to identify the patterns in general.

The proposed CNN contains layers of

- Layer 1: Conv1D (filters=1024, kernel_size=5, padding=" same")
- Layer 2: Dropout (0.2)
- Layer 3: Flatten ()
- Layer 4: Dense (8)

The CNN has "sparse categorical cross entropy" as the data is sparse and so it is used as loss function and "RMSprop " as activation function for adaptive learning rate.

3.3.3 DNN

Dense Neural Network (DNN) is a neural network where the layers are connected fully to each other. The proposed DNN contains layers as follows:

- Layer 1: Dense (1024)
- Layer 2: Dropout (0.2)
- Layer 3: Dense (8)

The DNN also has "sparse categorical cross entropy" as loss function and "RMSprop" as activation function.

3.3.4 LSTM

Long Short-Term Memory (LSTM) is used to work with a sequence of data. The proposed CNN contains the following layers:

- Layer 1: LSTM (1024)
- Layer 2: Dropout (0.2)
- Layer 3: Flatten ()
- Layer 4: Dense (8)

The performance of the above models was compared based on accuracy and f1 score.

3.4 Audio Transcription

With the negative and neutral audio obtained, the audio is converted into 25ms frames and fed to a transformer which is a self-supervised learning model, pre-trained with 100 hr audio and 53k unlabeled data. This Wav2Vec2 model processes the audio with the help of Librosa, a python module. After converting analog to digital, the features are passed to the Wav2Vec2 tokenizer which processes the audio to find the human speech. Then it is passed to the Wav2Vec model which generates the transcription present in the audio.

3.5 Binary Classiftcation

Text classifier models are trained for identification of suicidal intent from depression posts. For this, suicide depression detection dataset from Kaggle is used. The models BERT, ELECTRA and RoBERTa were trained and compared based on the accuracy and f1 score.

3.5.1 BERT

BERT is a transformer developed by engineers at Google. It works on two principles: Masked Language Model and Next Sentence Prediction. In MLM, almost 15% of the text is masked and BERT tries to guess the masked content. In Next Sentence Prediction, two sentences are provided and BERT guesses if they follow each other. By these encoder and decoder training, high performance can be achieved. For classification, BERT-base which is made of 12 layers of encoders and 768 hidden layers, is used for training and classification.

3.5.2 ELECTRA

ELECTRA is a transformer developed by engineers at Google. It is a combination of generators and discriminators. The training is done similar to BERT except that instead of masking, generators substitute the word and discriminators try to identify them. This allows us to train with much lesser computation power and can pit against BERT in performance. ELECTRA works like a combination of General Adhesive Network and BERT.

3.5.3 RoBERTa

RoBERTa stands for Robustly Optimised BERT using Pretraining Approach. This model is developed by engineers at META. The baseline of RoBERTa is BERT, which is optimised by much higher training. The RoBERTa model is essentially a BERT which was trained on larger sequences and larger batch sizes while removing the Next Sentence Prediction part. Also, instead of masking done

fixed for a sequence in BERT, dynamic masking is done for RoBERTa. For training, roberta-base model is used.

4 RESULTS AND DISCUSSION:

4.1 EVALUATION METRICS

To evaluate the models implemented, we use the Accuracy and F1 score of the models as accuracy provides an overview of the model's correctness and the F1 score provides an outlook of classification for each case as the cleaned dataset is not uniform between the classes. For metrics, we use the values (TP, TN, FP, FN) obtained from the confusion matrix which are described as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \ 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Table 1: Results of Emotion classification

Models	Accuracy	Precision	Recall	F1-Score
Random Forest	77	77	77	77
CNN Model	83	83	81	81
LSTM Model	84	83	84	84
DNN Model	**88**	**87**	**87**	**87**

Table 2: **Results of Text Classiftcation**

Models	Accuracy	Precision	Recall	F1-score
BERT	0.9694	0.9766	0.9608	0.9686
ELECTRA	0.9735	0.9687	0.9777	0.9732
ROBERTA	**0.9881**	**0.9882**	**0.9877**	**0.9879**

Table 3: Comparison Results

Comparison Study	Models	Accuracy	F1-Score	Precision	Recall
Our Model	**RoBERTa**	**98.81**	**98.82**	**98.77**	**98.79**
LR		63	46	35	78
Accoustic +	SVM	64	38	32	52
Linguistic Features [2]	RF	80	45	32	84
XGBOOST		77	47	37	74
DNN		77	3624	70	
DCM [8]	SVM	90	8482.5	85.5	

From Table 1, we can infer that DNN does well in classifying emotions based on MFCC features. For emotion classification, the proposed DNN outperforms the model mentioned in [2] by 6% in accuracy and f1-score. From Table 2, we can also infer that RoBERTa can outperform all the other mentioned transformers (BERT and ELECTRA), achieving a staggering 98% accuracy and F1 score. While compared to [1], the proposed model as a whole can outperform the suggested top models XGBoost and DNN in accuracy and F1 score. From Table 3, we can infer that the depression classification model which is a SVM on custom data [2] achieves 90% accuracy but the proposed RoBERTa performs better (98%) accuracy for detecting suicidal intent. This implies, the model can be used out in real world use cases.

5. CONCLUSION

Many people make online suicide notes as text or as audio which provides serious evidence, yet ignored by many. As the suicide rate is increasing, especially in countries like India and the US, the earlier detection of suicide intent,

especially among the youth can save the lives and possibly avert much loss for the country. This proposed method would be useful in detecting the user with suicidal intent at the earliest possible stage, automatically without any intervention required. After detecting intent, any form of help could be offered as soon as possible. The model could be implemented to work with real data in real time. With high accuracy, the prediction of people who intend to commit suicide and the signs ignored by others can be identified and they can be provided with proper counseling and care which can change the individual's life and nation's future.

6. FUTURE WORKS

Currently the model works well with English and can be made multi linguistic to support all people irrespective of the language they speak. Also, the model can be made to support unstructured data such as videos and images so that a single model can be used to analyze all types of unstructured data.

REFERENCES

[1] S. P. Dubagunta, B. Vlasenko and M. Magimai.-Doss,"Learning Voice Source Related Information for Depression Detection", *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019*, pp. 6525-6529, doi: 10.1109/ICASSP.2019.8683498.

[2] A Belouali, S Gupta, and V Sourirajan ,"Acoustic and language analysis of speech for suicidal ideation among US veterans", Journal on BioData Mining 14 (2021), pp. 11–11.

[3] S Scherer, J Pestian, and L.P Morency, "Investigating the speech characteristics of suicidal adolescents", *IEEE International Conference on Acoustics, Speech and Signal Processing* (2013).

[4] Sandeep Nallan Chakravarthula, Md Nasir, Shao-Yen, Tseng Haoqi, Li Tae Jin, Park Brian, Baucom Craig J. Bryan, Shrikanth Narayanan, Panayiotis Georgiou, ed. "Towards Real-Time Multimodal Emotion Recognition among Couples*", ICMI '20: Proceedings of the 2020 International Conference on Multimodal InteractionOctober*. 2020.

[5] Bhanusree Yalamanchili, Keerthana Dungala, Keerthi Mandapati, Mahitha Pillodi & Sumasree Reddy Vanga ,"Survey on Multimodal Emotion Recognition (MER) Systems*", Conference on Machine Learning Technologies and Applications and Algorithms for Intelligent Systems*. Singapore: Springer, 2021.

[6] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017*, pp. 2257-2260, doi: 10.1109/WiSPNET.2017.8300161

[7] J Briskilal and C N Subalalitha ,"An ensemble model for classifying idioms and literal texts using BERT and RoBERTa", *International Journal* (2022).

[8] Q. Huang, R. Chen, X. Zheng and Z. Dong, "Deep Sentiment Representation Based on CNN and LSTM," *2017 International Conference on Green Informatics (ICGI)*, 2017, pp. 30-33, doi: 10.1109/ICGI.2017.45.

[9] G. -M. Lin, M. Nagamine, S. -N. Yang, Y. -M. Tai, C. Lin and H. Sato, "Machine Learning Based Suicide Ideation Prediction for Military Personnel," in *IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 1907-1916, July 202*0, doi: 10.1109/JBHI.2020.2988393.

[10] S Arora and P Doshi,"A survey of inverse reinforcement learning: Challenges, methods and progress",*Artificial Intelligence* 297 (2021), pp. 103500–103500.

[11] T Miller ,"Explanation in artificial intelligence: Insights from the social sciences", *Artificial Intelligence* 267 (2019), pp. 1–38.

[12] A Muthumari and K Mala ,"Computerized Methods for Audio Segmentation and Classification: Survey", *International Journal of Applied Engineering Research* 10 (2015), pp. 26857–26870.

[13] Praveen Paul , "A Look into the Dark Pages of Usury by Ravenous Loan Sharks in India - A Review on Media Reports", *International Journal of Social Sciences* 9(1) (2020).

[14] Jinsong Su, Jialong Tang, Hui Jiang, Ziyao Lu, Yubin Ge, Linfeng Song, Deyi Xiong, Le Sun, Jiebo Luo ,"Enhanced aspect-

based sentiment analysis models with progressive self-supervised attention learning",*Artificial Intelligence* 296(2021).

[15] N Umakanth and S.Santhi , "CLASSIFICATION AND RANKING OF TRENDING TOPICS IN TWITTER USING TWEETS TEXT", Journal of Critical Reviews,2020, 7(04). https://doi.org/10.31838/jcr.07.04.171