International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

ML Based Approach for Speech Emotion Recognition

Prof. Ayesha A. Sayyad, Vishal Surya, Mahesh Chavhan, Subhash Lavand, Chaitanya Kisave

Abstract –The investigation of human discourse is a difficult examination territory as it concerns the discovery of user networks. Feelings assume an underlying part in human communication. The capacity to comprehend human feelings by breaking down voice is alluring in various uses of discourse acknowledgment in feelings can be found in various zones, for example, the association among PCs and people and call focuses. Already, feeling acknowledgment utilized straightforward classifiers on bag- of-words models. Nonetheless, the current work of feeling acknowledgment on Voice was done with the assistance of profound learning strategies on static voice information. The proposed strategy centers around expanding the general precision of feeling acknowledgment during calls utilizing man-made consciousness. The general point is to precisely perceive the different feelings that a specific speech communicates semantically.

.

Keywords- Human Emotion, feature extraction, machine learning, Speech conversion.

I. INTRODUCTION

The method by which a computer automatically detects human emotions and emotion-related states from speech is known as speech emotion recognition (SER). Human intellect, reasoned decision-making, social interaction, perception, memory, learning, and creation are all significantly influenced by emotion. The ability to transmit emotions is a key characteristic that sets humans apart from other animals as a higher species. There are several real-world applications for speech emotion identification, including online learning environments, contact centers, and depression diagnosis. The neural network approach has swept numerous fields as deep learning flourishes. Deep learning has fueled significant advancements in the field of speech emotion recognition, leading to significant performance improvements.

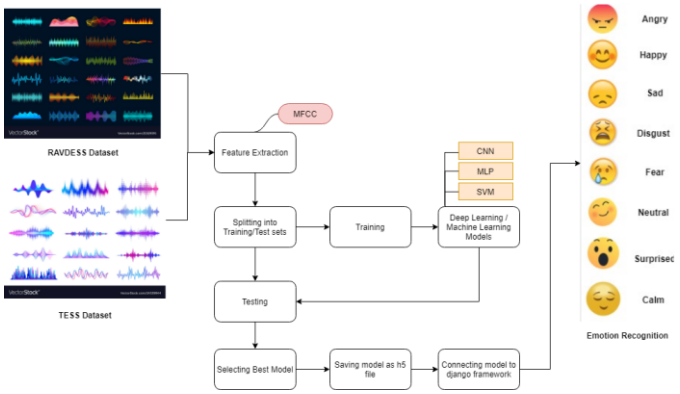
Additionally, the primary network architecture was specifically created to address issues in various domains. One of the main issues in speech emotion detection is how to use the network in other domains in a way that makes sense in order to enhance the capacity to model emotional information. Furthermore, the recognition task is more difficult because to the lack of datasets and emotion perception. As such, speech emotion recognition performance is still not optimal. The study of the composition and operation of the many brain regions in humans that are responsible for emotion, cognition, and awareness has gotten stronger in recent years because to advancements in brain science.

II. IMPLEMENTATION

Text perceives from human discourse utilizing discourse transformation library through component extraction strategies. Human Mood States is a mental instrument for surveying the person's temperament state. It characterizes 65 descriptive words that are appraised by the subject on the five-point scale. Every descriptive word adds to one of the four classifications. The higher the score for the modifier, the more it adds to the general score for its classification, aside from loose and effective

whose commitments to their individual classifications are negative. Disposition states consolidates these evaluations into a four-dimensional temperament state portrayal comprising of classifications: outrage, glad, tragic and typical. Contrasting with the first structure, we disposed of the modifier blue, since it just seldom relates to a feeling.

A. **System Architecture**



B. **Algorithm:**

1. LSTM :

Since a typical RNN just has one hidden state that is transferred across time, learning long-term dependencies may be challenging for the network. In order to solve this issue, LSTMs introduce memory cells, which are long-term information storage units. Because LSTM networks can learn long-term relationships from sequential data, they are a good fit for applications like time series forecasting, speech recognition, and language translation. For the study of images and videos, LSTMs can also be used in conjunction with other neural network architectures, such as Convolutional Neural Networks (CNNs).

Three gates—the input gate, the forget gate, and the output gate—control the memory cell.

The information that is added to, subtracted from, and output from the memory cell is determined by these gates. What data is added to the memory cell is managed by the input gate. What data is erased from the memory cell is managed by the forget gate. Additionally, the output gate regulates the data that the memory cell outputs. Because of this, long-term dependencies can be learned by LSTM networks by allowing them to choose keep or reject information as it passes through the network.

Forget Gate

The forget gate eliminates data that is no longer needed in the cell state. The gate receives two inputs, xt (the input at that specific moment) and ht-1 (the output of the previous cell), which are multiplied by weight matrices before bias is added. After being run through an activation function, the output is binary. When the output for a certain cell state is 0, the information is lost, and when the output is 1, it is saved for later use.

The forget gate's equation is as follows:

$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ where

The weight matrix connected to the forget gate is denoted by W_f.

The concatenation of the current input and the prior hidden state is shown by the notation [h_t-1, x_t].

The bias with the forget gate is b_f.

The sigmoid activation function is denoted by σ.

Input gate

The input gate modifies the cell state by adding pertinent information. Using inputs ht-1 and xt, the sigmoid function is first used to regulate the information before filtering the values to be remembered in a manner akin to a forget gate.

Output gate

The output gate's job is to extract valuable information from the current cell state so that it can be shown as output. The tanh function is first used to the cell in order to build a vector. The data is then filtered by the values that need to be remembered using inputs ht-1 and xt, and the sigmoid function is used to regulate the data.

2. CNN:

One kind of deep learning algorithm that works especially well for tasks involving picture recognition and processing is the convolutional neural network (CNN). Convolutional, pooling, and fully connected layers are some of the layers that make it up. The human brain's visual processing served as the inspiration for CNN architecture, which makes them ideal for identifying spatial connections and hierarchical patterns in images.

Convolutional Layers: These layers use filters, sometimes referred to as kernels, to apply convolutional operations on input pictures in order to detect features like textures, edges, and more intricate patterns. The spatial associations between pixels are preserved with the use of convolutional techniques.

Pooling Layers: By downsampling the input's spatial dimensions, pooling layers lower the network's computational complexity and parameter count. A typical pooling operation called "max pooling" chooses the largest value among a set of adjacent pixels.

Functions of Activation: Rectified Linear Unit (ReLU) and other non-linear activation functions add non-linearity to the model, enabling it to discover more intricate links in the data.

Fully Connected Layers: These layers use the high-level features that the preceding layers have learned to make predictions. They link each

3. MMLDA Algorithm for summarization

Steps:

1. For the topic T, draw $\varphi^{TG} \sim Dir(\lambda^{TG})$ and $\varphi^{VG} \sim Dir(\lambda^{VG})$ denote the general textual distribution and visual distribution, respectively. $Dir(\cdot)$ is the Dirichlet distribution. Then draw $\phi^{Z} \sim Dir(\beta^{Z})$, which indicates the distribution of subtopics over the microblog collection corresponding to T.
2. For each subtopic, draw $\varphi_K^{TS} \sim Dir(\lambda^{TS})$ and $\varphi_K^{VS} \sim Dir(\lambda^{VS})$, $k \in \{1, 2, \ldots, K\}$, correspond to the specific textual distribution and visual distribution.
3. For each microblog M_i, draw $Z_i \sim Multi(\phi^{Z})$, corresponds to the subtopic assignment for Mi. Multi(·) denotes the Multinomial distribution. Then draw $\phi_i^R \sim Dir(\beta^R)$ indicates the general-specific textual word distribution of M_i. Similarly, draw $\phi_i^Q \sim Dir(\beta^Q)$ indicates that for visual words.
4. For each textual word position of M_i, draw a variable $R_{ij} \sim Multi(\phi_i^R)$:
   - If R_{ij} indicates General, then draw a word $W_{ij} \sim Multi(\varphi^{TG})$.
   - If R_{ij} indicates Specific, draw a word W_{ij} from the Z_i-th specific distribution $W_{ij} \sim Multi(\varphi_{Z_i}^{TS})$
5. The generation of visual words is similarly done as in step 4.

III. RESULTS AND DISCUSSION

The experimental result evaluation, we have notation as follows:

TP: True positive (correctly predicted number of instance)

FP: False positive (incorrectly predicted number of instance),

TN: True negative (correctly predicted the number of instances as not required)

FN false negative (incorrectly predicted the number of instances as not required),

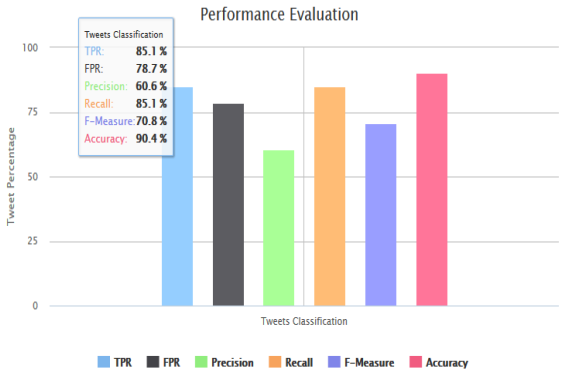
On the basis of this parameter, we can calculate four measurements

Accuracy = TP+TN÷TP+FP+TN+FN

Precision = TP ÷TP+FP

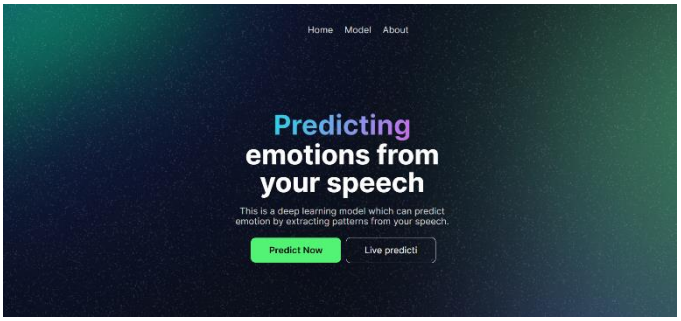
Recall= TP÷TP+FN

F1-Measure = 2×Precision×Recall ÷Precision+ Recall.

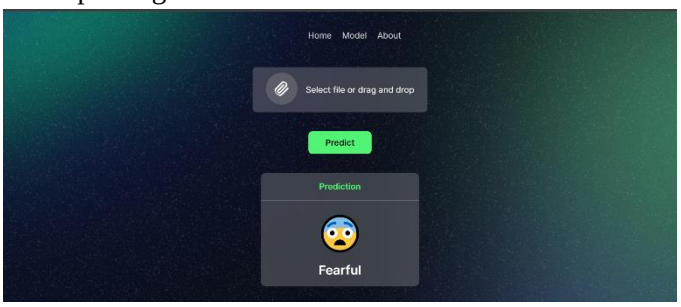


Parameters	Percentage
TPR	85.1
FPR	78.7
Precision	60.6
Recall	85.1
F-Measure	78.8
Accuracy	94.4

3.Home Page:

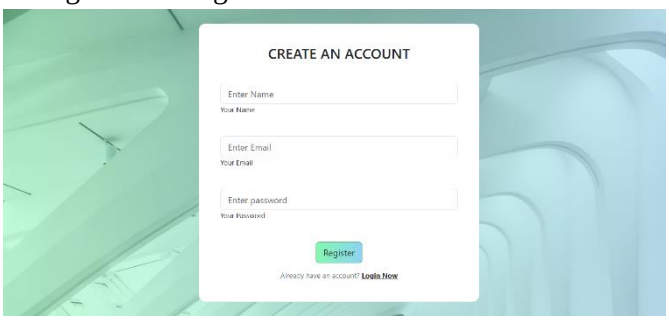


4. Output Page:

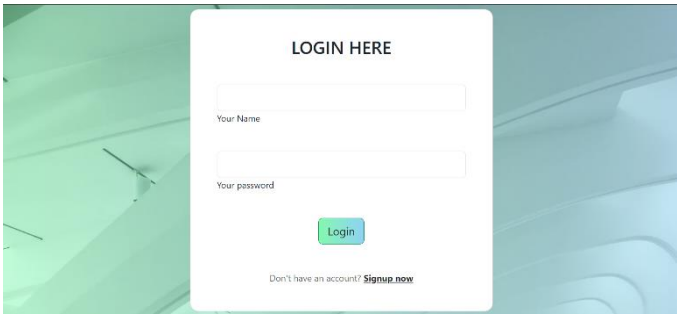


V. OUTPUT

1. Registration Page:



2. Login Page



VI. Conclusion

The method of recognizing the temperament or state of an individual through voice is an arising thought where the helpfulness of this cycle is inescapable, and will impart its uses to numerous areas from clinical to data advancements. This venture actualizes an Emotion Recognition on Speech utilizing novel technique a Profile of Mood States (POMS)using multinomial innocent Bayes speaks to four-dimensional temperament state portrayal utilizing 65 modifiers with blend of feelings classifications like upbeat, miserable, outrage and typical.

REFERENCES

[1] K.Tarunika , R.B Pradeeba , P.Aruna" Applying Machine Learning Techniques for Speech Emotion Recognition" ICCCNT 2018.

[2] Surekha Reddy B, T. Kishore Kumar" Emotion Recognition of Stressed Speech using Teager Energy and Linear Prediction Features" 2018 IEEE 18th International Conference on Advanced Learning Technologies

[3] Li Zheng, Qiao Li2, Hua Ban , Shuhua Liu1" Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest"IEEE 2018

[4] O. Irsoy and C. Cardie, "Opinion Mining with Deep Recurrent Neural Networks," in Proc. of the Conf. on Empirical Methods in Natural Language Processing. ACL, 2014, pp. 720–728.

[5] S. M. Mohammad and S. Kiritchenko, "Using Hashtags to Capture Fine Emotion Categories from Tweets," Computational Intelligence, vol. 31, no. 2, pp. 301–326, 2015.

[6] B. Nejat, G. Carenini, and R. Ng, "Exploring Joint Neural Model for Sentence Level Discourse Parsing and Sentiment Analysis," Proc. of the SIGDIAL 2017 Conf., no. August, pp. 289–298, 2017.

[7] Michael Neumann, Ngoc Thang Vu," IMPROVING SPEECH EMOTION RECOGNITION WITH UNSUPERV"IEEE 2019.

[8] Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller "END-TO-END SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS"IEEE 2018

[9] Po-Wei Hsiao and Chia-Ping Chen" EFFECTIVE ATTENTION MECHANISM IN DYNAMIC MODELS FOR SPEECH EMOTION RECOGNITION"IEEE 2018

[10] Saikat Basu, Jaybrata Chakraborty, Md. Aftabuddin" Emotion Recognition from Speech using Convolutional Neural Network with Recurrent Neural Network Architecture" International Conference on Communication and Electronics Systems (ICCES 2017)