International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7938
***Article Received**: 05 May 2023 **Revised**: 26 June 2023 **Accepted**: 15 July 2023*

Integrated Approach for Emotion Detection via Speech and Text Analysis

Dipika Birari[1], Gajanan Walunjkar[2], Aarti Dandavate[3], Sonali Mallinath Antad[4], Sheetal Phatangare[5]

[1]Assistant Professor, Department of Information Technology, Army Institute of Technology, Savitribai Phule Pune University, Pune, Maharashtra, dipikabirari001@gmail.com

[2]Assistant Professor, Department of Information Technology, Army Institute of Technology, Savitribai Phule Pune University, Pune, Maharashtra, gwalunjkar@aitpune.edu.in

[3] Associate Professor, Department of Computer Engineering, Dhole Patil College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, dr.aratid@dpcoepune.edu.in

[4]Assistant Professor, Department of Computer Engineering, Vishwakarma Institute of Technology, Savitribai Phule Pune University, Pune, Maharashtra, sonalitangi@gmail.com

[5]Assistant Professor, Department of Computer Engineering, Vishwakarma Institute of Technology, Savitribai Phule Pune University, Pune, Maharashtra, phatangaresheetal@gmail.com

Abstract: This paper aims to provide a comprehensive solution for effective reviews using deep learning models. Customers often have difficulty to find accurate reviews of the things they are interested in. The proposed framework implements a review mechanism to address this problem, which will give customers relevant reviews based on video reviews supplied in the product description. The goal of this system is to turn video reviews into a particular rating so that viewers may get a summary of the review without having to watch the full thing by simply glancing at the rating. Deep learning neural networks are used by the model for both text and audio processing in order to achieve this. The well-known RAVDESS dataset serves as the basis for the audio model's training and offers a wide range of emotional expressions. The suggested system uses two methods to provide reviews: text-based natural language processing and audio frequency spectrograms. By utilizing these two techniques, it may provide consumers accurate and trustworthy ratings while guaranteeing that the review procedure is not impeded. The aim is achieved with high accuracy to ensure that users can make informed decisions when purchasing products based on the provided reviews. With the aid of this review system, customers will be able to quickly find out crucial details about a product they are interested in, thus increasing their pleasure and loyalty.

Keywords: Spectrogram, Speech emotion recognition, NLP, convolutional neural network.

I. Introduction

Humans use a variety of techniques, such as verbal and nonverbal communication, to convey their emotions. One of the most common ways to express emotions, including happiness, sadness, anger, and excitement, is through speech. Due to the ease with which others may infer a speaker's emotions from the tone and pitch of their voice, the way these emotions are expressed verbally may be a good indicator of their nature. For example, a high pitch is typically associated with surprise, while a low pitch is associated with sadness. This ability to recognize emotions is due to the biological neural network in the human brain. Users who make purchases from e-commerce websites must go through a drawn-out and tedious procedure of filling out multiple forms and checking boxes. Why not automate this process so that all one needs to do is record a video of themselves giving a product a review, and the feedback would be produced automatically along with a brief description of keywords? Users may also submit text reviews, which can be used to rate and offer comments.

The Neural Networks and Machine Learning can be utilized to develop a system that can identify emotions in speech. Despite the abundance of research proposals, developing an effective approach for understanding speech emotions remains a significant challenge. A spectrogram is a visual representation of the amplitude or strength of a signal at various frequencies over time. It provides an illustration of how the energy levels of a waveform vary over time, allowing one to determine whether the energy at a given moment is higher or lower than at other moments. By examining the spectrogram, one can gain insight into the changes in energy levels over time and at different frequencies.

The goal of the suggested framework is to translate a video review into a particular rating so that a user may read the list of reviews by simply looking at their rating without having to watch the entire thing. The framework is as shown below in figure 1.

***Article Received**: 05 May 2023 **Revised**: 26 June 2023 **Accepted**: 15 July 2023*

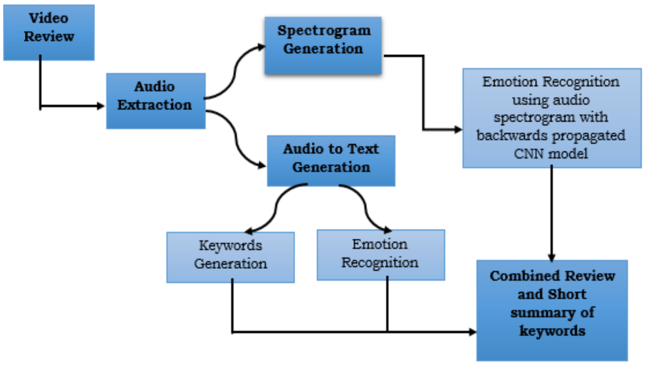

Video Review
Audio Extraction
Spectrogram Generation
Audio to Text Generation
Emotion Recognition using audio spectrogram with backwards propagated CNN model
Keywords Generation
Emotion Recognition
Combined Review and Short summary of keywords

Fig. 1. System Architecture

Stevens, Volkmann, and Newmann developed a unit of pitch in [11], so that listeners would perceive equivalent lengths in pitch as equal size. The mel-scale is the name for this. The frequency of a spectrogram are translated to the mel- scale to create a mel- spectrogram. For the system's audio and text processing, deep learning neural networks are used. The RAVDESS dataset for audio files is used for training. The objective is to provide the finest evaluation of a product through its video. Along with being turned into audio, the video review is also transformed into text so that Natural Language Processing can provide the necessary review rating.

- Natural Language Processing (NLP): It is an area of artificial intelligence that is concerned with teaching computers to comprehend and analyse human language. To analyse, interpret, and produce natural language text and voice on computers, a variety of methods and algorithms are used in NLP. It involves activities like text classification, information extraction, speech recognition, language translation, sentiment analysis, and sentiment analysis.
- Convolutional Neural Network (CNN), which is a kind of deep learning neural network frequently employed in areas like natural language processing, picture and video identification, and others when the input has a grid-like structure. Through the use of several layers of filters and pooling processes, CNNs are built to automatically recognise patterns and features in incoming data, enabling them to efficiently learn hierarchical representations of the data.

II. Literature Survey

In recent years, the field of sentiment recognition from audio has seen several notable advancements. Earlier efforts focused on using a single spectrogram to train the model. However, with advancements in technology and computing power, it is now possible to process audio into multiple spectrograms. This has allowed for more precise analysis of the signal, leading to improved accuracy in sentiment recognition.

Shiqing Zhang et al. [1] utilized mel-spectrograms to differentiate between speeches, which were then fed into a pretrained AlexNet DCNN model for prediction. The results showed an 86.30% UAR accuracy on the EMO-DB dataset, surpassing the three other works that were compared, with accuracies of 79.1%, 84.6%, 86.0%, and 86.1%.

M.E. Seknedy et al. [2] trained four different machine learning models using a variety of voice feature combinations. The models used were Multi-Layer Perceptron, Logistic Regression, Support Vector Machine, and Random Forest. The feature set included the major speech properties of prosody, spectrality, and energy. The proposed model utilized cross-corpus data from multiple languages to recognize voice emotions. The results showed an accuracy of 82.22% for the single corpus, and for cross-corpus analysis, the accuracy for RADVESS, EmoDB, and CaFE was 79.26%, 88.24%, and 82.35%, respectively.

In paper [3] JORGE OLIVEIRA et al. dataset's features were extracted and crossfold validation was conducted. The resulting dataset was then fed into a DNN to construct a neural network. The network's weights were established after 50 epochs, with emphasis placed on selecting the weights with the lowest magnitude. Upon testing, the results demonstrated an F1 accuracy score of approximately 68%, surpassing the accuracy of previous models for the MELD data.

Sabrine Dhaouadi et al. [4] the classification of opposing emotions was achieved through the use of Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs). Human speech can be analyzed to extract various temporal and spectral characteristics. Both SVM and ANN are commonly used for identifying emotions based on speech signals. A comparative analysis was conducted to evaluate the performance of these two methods. The findings suggest that the ANN model outperformed the SVM model under the fixed parameters used in the experiments.

M. Gokilavani et al. [5] utilized Convolutional Neural Network Engineering(CNN) in this application to analyze and categorize audio data while implementing emotion identification, and TensorFlow is employed as the backend. Convnet oper- ates similarly to human brain neurons, and it also reduces error rates by 5 to 10%.The three datasets employed in this study had very high accuracy: the Ravdness dataset had a 96% accuracy rate, the Tess dataset had a 99% accuracy rate, and the Crema-D dataset had an accuracy rate of 84%.

X. Liu et al. [6] the sliding window approach was utilized to extract window-level features, which were then merged with the interval-level feature for training the ANN. The window-level features were obtained by analyzing the speech signal using a sliding window, and the interval-level feature was calculated separately. The MFCC pre-trained model was used in combination with ANN, resulting in a high accuracy and f1-score of approximately 91%, precision of 84%, and recall of 99%.

In paper [7], SUDARSANA REDDY KADIRI et al. divide speech into two categories: neutral and emotional, and compute the Kullback-Leibler (KL) distance to assess how closely the feature distribution of these two categories resembles one another. The system chooses the best detection technique based on the KL distance value. The fact that this technique does not require emotive speech to be utilised for training is a key benefit.

REZA LOTFIAN et al [8] natural speech was transformed into synthetic speech using TTS generators. The artificial speech generated by ten different TTS generators was fed into a conventional neural network. The results indicate better accuracy compared to other models that have been previously tested.

S. Prasomphan [9] in comparison to previous approaches, the proposed model simplifies the classification process and reduces the number of features required for the neural network. This is achieved by extracting the essential features from the spectrogram and utilizing a novel network architecture. The results demonstrate that the accuracy for anger in the Berlin dataset was 80.67%, for happiness 82.47%, and for fear 83.28%.

According to Kun-Yi Huang et al. [10], a prosodic phrase auto-tagger was utilised to extract the verbal/nonverbal segments from an SVM-based verbal/nonverbal sound detector. Convolutional neural networks were utilized to extract emotion and sound characteristics for each segment, and these were combined to form a general feature vector based on CNN. Most existing emotion identification algorithms focus solely on a limited range of nonverbal noises that are typically present in everyday speech. In contrast, this study considered both verbal and nonverbal sounds within an utterance for emotion recognition in realistic conversational contexts.

III. System Implementation

System Implementation contains Generation of audio from video, Generation of keywords, Extraction of an emotion from audio, Emotions Extraction from transcript as shown in figure 2.

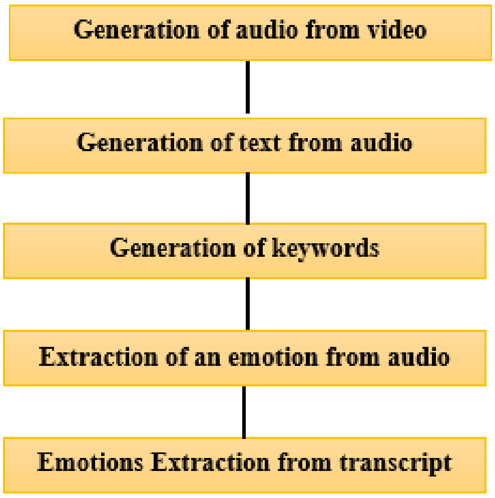

Generation of audio from video
Generation of text from audio
Generation of keywords
Extraction of an emotion from audio
Emotions Extraction from transcript

Fig 2: Phases of system implementation

A. Generating audio from video

generating audio from video which can be done using the python library MoviePy. The extracted audio is then passed on to our text generation model. Folowing are the steps for To extract audio from a video using Python MoviePy library.

- Step 1: Install the moviepy library using pip. One can do this by running pip install moviepy in terminal or command prompt.
- Step 2: Import the necessary modules in Python script like VideFileClip.
- Step 3: Load the video file using VideoFileClip() method.
- Step 4: Pull out the audio from the video using the audio property of the video object.
- Step 5: Save the audio as a file using the write_audiofile() method.

B. Generating text from audio

Any valid audio format may be used as the input for this model's audio file. The model only takes audio input as.wav files, hence the audio file that was received as input has to be converted into the WAV(.wav) format. With the aid of the AudioSegment module from the PyDub Python library, this is possible.

The text extraction model can be fed with this transformed input audio file. It uses the Wav2vec model, which was developed using data from the Librispeech dataset, which includes audio samples that were captured at a sampling rate of 16 kHz. The model creates a transcript of the specified audio input, which is essentially the audio file's content but in written form.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7938
***Article Received**: 05 May 2023 **Revised**: 26 June 2023 **Accepted**: 15 July 2023*

Algorithm: Text Generation from Audio

```
1. Install required libraries and packages
2. Since the model take only wav file as input, convert the audio format to wav using pydub.
3. Split the audio in 3 minutes each to minimize the computational power used by model.
4. Check the file generated in audio directory
5. Load model
6. Load tokenizer
7. Pass the audio as input to the model
8. Print the output
```

C. Generation of keywords

The spacy Python package is used to extract the keywords from the transcript audio. Spacy, a Python NLP library that is simple to use and appropriate for beginners, is comparable to BERT. The text produced in the prior stage may be used to extract the keywords using the en_core_sci_scibert pipeline.

D. Extracting emotion from audio

The pretrained Resnet-18 CNN is employed to extract emotion from audio. Convolutional neural network ResNet-18 has 18 layers and trained using data from the ImageNet database and over one million photos. A Resnet-18 CNN's fundamental structure is as shown in fig. 3

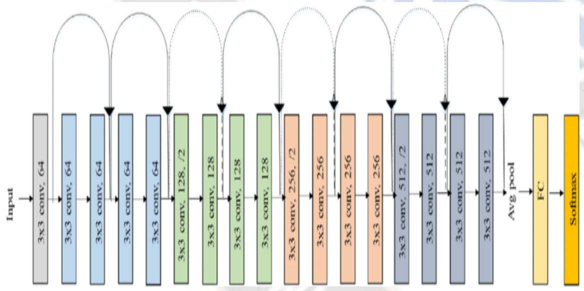


Fig 3: Fundamental structure of Resnet-18 CNN

Replace the first and last layers of the CNN depending to input and output while designing the Resnet-18 model. The last layer is replaced with a sequential layer to match the numerous output classes that are needed for output, and the first layer is replaced by a convolution layer inside channels according to the spectrogram that will pass.

E. Extracting emotions from transcript

Extracting emotions from a transcript involves identifying the emotional state of the speaker based on their choice of words, tone of voice, and other nonverbal cues. Here are the steps to be followed to extract emotions from the transcript generated:

1. **Analyze the transcript carefully**: Start by analyzing the transcript from start to finish to get an overall sense of the conversation.
2. **Recognize emotive terms:** Look for emotive words like "happy", "sad", "angry", "frustrated", "excited", and so on.
3. **Consider the context**: Remember that emotions are often expressed in the context of the conversation. Consider the overall context of the conversation to get a better understanding of the emotional state of the speaker.

This module's emphasis is on the in-depth analysis of the text; which generated in the previous module. The main goal is to do a thorough sentiment analysis of the language, which may be employed to derive important information about how the consumer views the product. To clearly comprehend the user's feelings towards the product, this analysis will be carried out using the audio transcript. By conducting this study, one might hope to get a greater comprehension of the user's experience and degree of product satisfaction.

This module use a broad dataset made up of user reviews gathered from a number of well-known websites, including Yelp, IMDB, and Amazon. Users may express their ideas and post reviews on a variety of goods and services on these websites. The dataset consists of 2742 reviews from several sources, including 998 reviews from Yelp, 746 reviews from IMDB, and 998 reviews from Amazon. Each review in the dataset is labeled with a target value of either 0 or 1, indicating whether the sentiment expressed in the review is positive or negative, respectively. By leveraging this comprehensive dataset, the goal is to conduct an extensive sentiment analysis of the reviews to gain valuable insights into the users' perception and experience of the products and services. In believe that this analysis will provide with a deeper understanding of the users' sentiments, which can be leveraged to improve the quality of the products and services offered.

The performance of this sentiment analysis model is assessed using the two key metrics, the F1-score and Matthews Correlation Coefficient (MCC). While MCC assesses the accuracy of binary classifications by accounting for true and false positives and negatives, the F1-score is a widely used assessment tool that provides a balance between precision and recall. These metrics may be used to evaluate the model's efficiency and accuracy in properly predicting the sentiment of the reviews. The model may be modified as needed to improve performance using the insights provided by these indicators. The F1-score of a classification model is calculated as given in equation (1).

$$F1_{score} = {}^{2(P_r\ *R_e)}/_{P_r\ + R_e} \quad .....(1)$$

Where, Pr is the Precision and Re of Recall of the classification model.

***Article Received**: 05 May 2023 **Revised**: 26 June 2023 **Accepted**: 15 July 2023*

Following approaches used to analyze the result and have a greater insight of data.

1. Exploratory Data Analysis (EDA): In this approach generates multiple graphs and plots to understand the data by utilizing visual techniques to analyze data, enabling the identification of trends, patterns, and the verification of assumptions through the use of statistical summaries and graphical representations. For e.g. How many reviews have the positive reviews and how many have the negative reviews in the dataset?
2. Bidirectional Encoder Representations from Transformers (BERT): BERT (Bidirectional Encoder Representations from Transformers) is a machine learning framework that is open-source and commonly used for various natural language processing (NLP) tasks. In contrast to other methods that employed either left-to-right or combined left-to-right and right-to-left training approaches for text sequences, BERT introduces bidirectional training. This unique training methodology allows the model to gain a more comprehensive understanding of language context and coherence compared to single-direction language models. By considering information from both preceding and subsequent words, BERT can capture a richer understanding of the relationships and dependencies within a given text.

   To establish context with reviews, one can fine-tuned BERT using required dataset. This involved training BERT on our specific data to better capture the nuances and patterns in the review context. For evaluating the performance of this approach, calculated relevant metrics used to measure the effectiveness and quality of the model's predictions.

3. A Robustly Optimized BERT Pretraining Approach (RoBERTa) : Similar to BERT, RoBERTa is a transformer-based language model that processes input sequences and produces contextualised word representations via self-attention mechanisms. RoBERTa distinguishes itself from BERT through its training methodology. RoBERTa was trained on a significantly larger dataset and implemented an enhanced training procedure. In addition to the larger dataset, RoBERTa incorporates a dynamic masking technique during training. In evaluating proposed approach using RoBERTa, the calculated relevant metrics used to assess the model's performance and effectiveness in the specific task or application.
4. XLNet: XLNet is a pre-trained language model with a distinctive feature known as permutation-based training, which enables XLNet to model the probability of word sequences without relying on a specific order of words. Its innovative training methodology and competitive performance have made it a valuable tool in the field of natural language processing.
5. ElLECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately(ELECTRA) is a pre-training approach for language models. It differentiates between relevant and irrelevant information within the input sequence, enhancing its ability to generate accurate and meaningful responses. Its effectiveness and versatility make it a valuable tool in the field of natural language processing.
6. DistilRoBERTa: It is a distilled version of the RoBERTa model, Distillation is a technique that involves training a smaller, faster, and less complex model to mimic the behavior of a larger and more complex model. It is used to achieve high-quality results with fewer computational resources and lower costs.

IV. Evaluation of Models

After fine-tuning the pretrained CNN's layers, it is trained and evaluated on the Ravdess dataset using spectrogram, mel-spectrogram, and mfcc inputs. Figure 4, figure 5 and figure 6 shows the metrics of the model for spectrograms, MFCC and Mel-spectogram respectively..

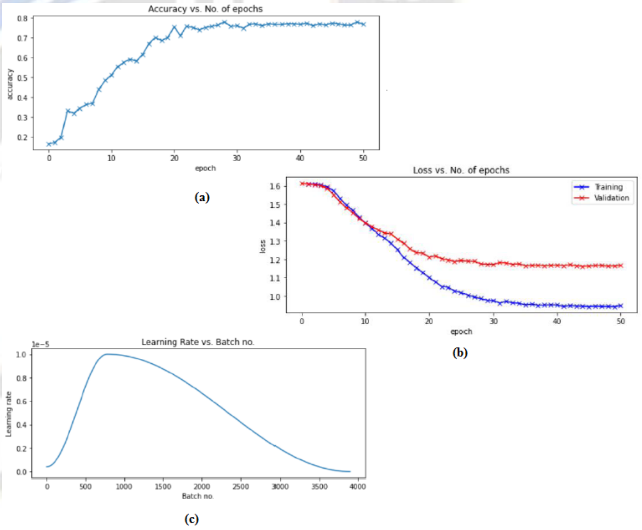


Fig 4: Spectrogram: (a) Accuracy vs No. of epochs (b) Loss vs No. of epochs (c) Learning Rate vs Batch no.

***Article Received**: 05 May 2023 **Revised**: 26 June 2023 **Accepted**: 15 July 2023*

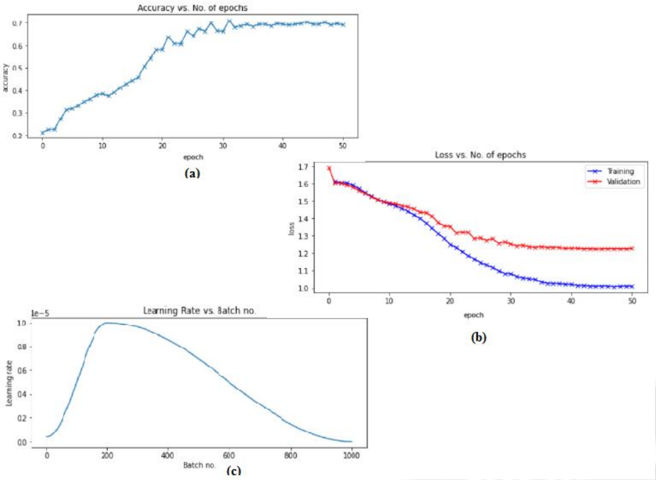


Fig 5: MFCC: (a) Accuracy vs No. of epochs (b) Loss vs No. of epochs (c) Learning Rate vs Batch no.

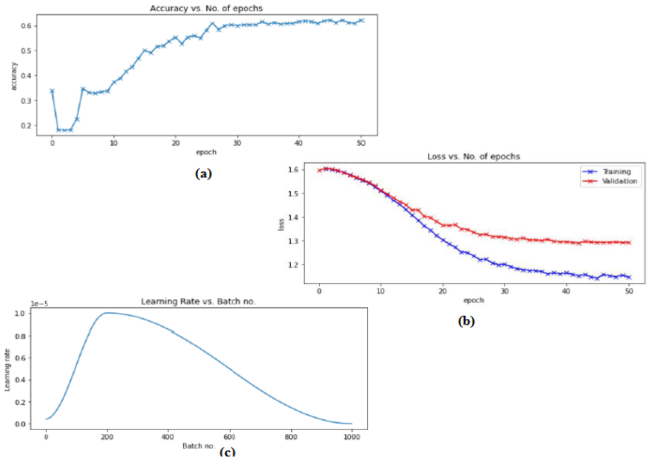


Fig 6: Mel-Spectrogram (a) Accuracy vs No. of epochs (b) Loss vs No. of epochs (c) Learning Rate vs Batch no.

The accuracy for the spectrogram model (76.5%) has higher than the mfcc and mel-spectrogram models, hence the spectrogram model is used for predicting the emotion from audio. The metric function "accuracy" is used is to evaluate the performance this model and accuracy is calculated as shown in equation 2,

$$\frac{T\ Negative + T\ Positive}{T\ Positive + F\ Positive + T\ Negative + F\ Negative} \qquad \text{.........(2)}$$

The training of the models generated the following accuracy metrics: F1 score and MCC (Matthews Correlation Coefficient). By examining these accuracy metrics, it helps to gain valuable insights into the models' performance. The F1 score offers a comprehensive evaluation by considering both precision and recall, while MCC provides a correlation-based measure that takes into account all classification outcomes. These metrics help in determining the models' effectiveness and suitability for specific tasks, enabling us to make informed decisions regarding their implementation and potential improvements. Comparison of Trained Models as shown in Table I.

Table I. Comparison of Trained Models

Model	F1-Score	MCC Score
BERT	0.9413919413919	0.8838209637966
RoBERTa	0.9694793536804	0.9381139608217
XLNet	0.955277280858	0.9090535425547
ELECTRA	0.9596969696969	0.9283550369286
DistilRoBERTa	0.9498207885304	0.8980697492424

V. Improvements Made

The proposed work implemented a process known as fine-tuning to enhance the performance of pretrained models such as BERT, RoBERTa, XLNet, DistilRoBERTa, and Electra. These models undergo initial training on large-scale datasets using advanced techniques, where the weights of their neurons are assigned specific values based on patterns and information in the training data.

However, given the uniqueness of each dataset, further refinement of these pretrained models becomes necessary to adapt them to the specific domain or task. To achieve this, a dataset specific to the project's requirements was utilized. This dataset comprised relevant examples and instances tailored to the problem at hand.

By training the pretrained models on the project's dataset, the weights were fine-tuned to better suit the specific requirements. This process involved exposing the models to the dataset, allowing them to learn and adjust their internal representations to capture the nuances and patterns inherent in the data.

VI. Experimental Results

The experiment results phases in terms of working modules are shown in figure 7.

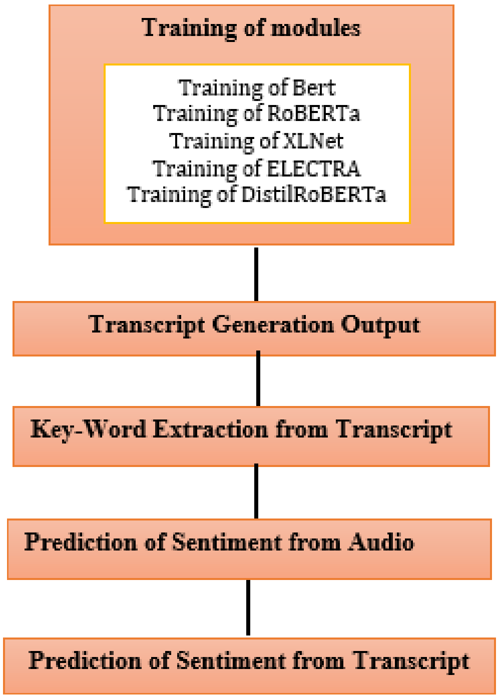

Training of modules
Training of Bert
Training of RoBERTa
Training of XLNet
Training of ELECTRA
Training of DistilRoBERTa
Transcript Generation Output
Key-Word Extraction from Transcript
Prediction of Sentiment from Audio
Prediction of Sentiment from Transcript

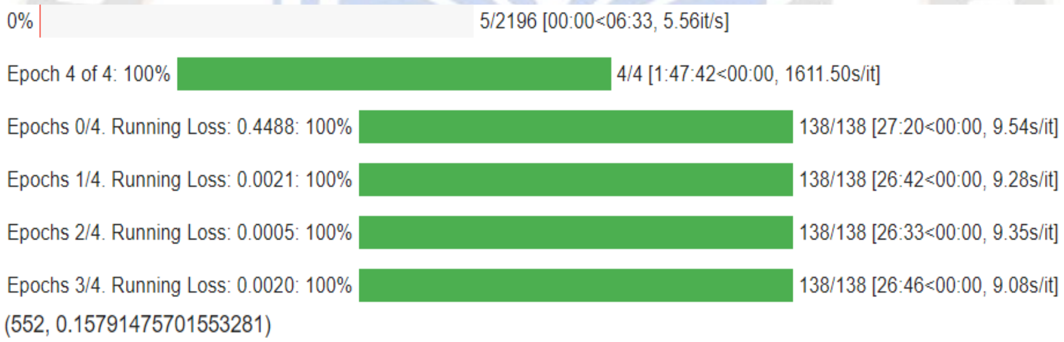
Fig 7: Phases of working module

A. Training of modules

1. Training of Bert – Training a BERT model in Figure 8 involves two main steps:

- **Pre-training**: During pre-training, BERT learns to predict missing words in a sentence using a large corpus of unlabeled text. The model is trained in a self-supervised manner, meaning it doesn't require labeled data for this step. BERT utilizes a transformer architecture, which is a type of deep neural network that allows for efficient parallel processing of sequential data.

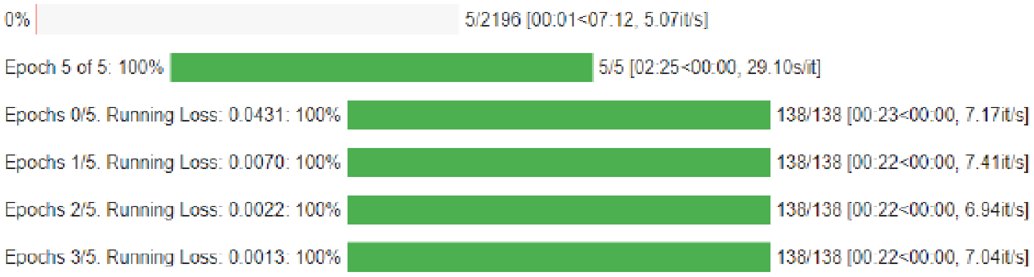
- **Fine-tuning:** After pre-training, the BERT model is fine-tuned on specific downstream tasks that require labeled data. Fine-tuning involves training the pre-trained BERT model on a smaller labeled dataset related to the specific task, such as sentiment analysis, named entity recognition, or question answering. During fine-tuning, additional task-specific layers are added to the pre-trained BERT model, and the entire model is trained on the labeled data.



```
0%                                       5/2196 [00:00<06:33, 5.56it/s]
Epoch 4 of 4: 100%                       4/4 [1:47:42<00:00, 1611.50s/it]
Epochs 0/4. Running Loss: 0.4488: 100%   138/138 [27:20<00:00, 9.54s/it]
Epochs 1/4. Running Loss: 0.0021: 100%   138/138 [26:42<00:00, 9.28s/it]
Epochs 2/4. Running Loss: 0.0005: 100%   138/138 [26:33<00:00, 9.35s/it]
Epochs 3/4. Running Loss: 0.0020: 100%   138/138 [26:46<00:00, 9.08s/it]
(552, 0.15791475701553281)
```

Fig 8: Training of BERT

2. Training of RoBERTa – RoBERTa has demonstrated significant improvements over BERT on various benchmark tasks in natural language understanding and generation. It achieves state-of-the-art results on tasks such as question answering, natural language inference, and sentiment analysis. Its enhanced pre-training methodology, larger training corpus, dynamic masking, and longer training duration contribute to its robustness and improved performance as shown in Figure 9 the training of RoBERTa.

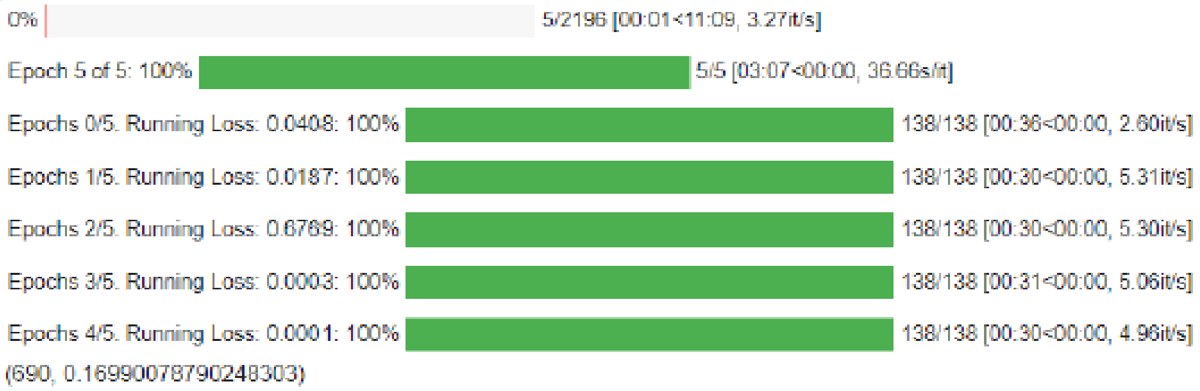


```
0%                                       5/2196 [00:01<07:12, 5.07it/s]
Epoch 5 of 5: 100%                       5/5 [02:25<00:00, 29.10s/it]
Epochs 0/5. Running Loss: 0.0431: 100%   138/138 [00:23<00:00, 7.17it/s]
Epochs 1/5. Running Loss: 0.0070: 100%   138/138 [00:22<00:00, 7.41it/s]
Epochs 2/5. Running Loss: 0.0022: 100%   138/138 [00:22<00:00, 6.94it/s]
Epochs 3/5. Running Loss: 0.0013: 100%   138/138 [00:22<00:00, 7.04it/s]
```

Fig 9: Training of RoBERTa

***Article Received**: 05 May 2023 **Revised**: 26 June 2023 **Accepted**: 15 July 2023*

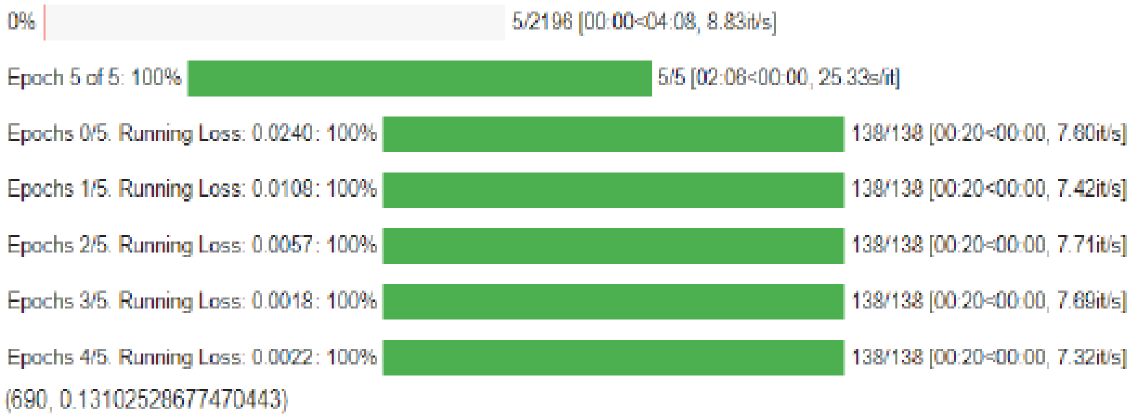
3. Training of XLNet – XLNet captures bidirectional context effectively by considering all possible permutations of the input sequence during training. With its large-scale training on unlabeled data and fine-tuning on specific tasks, XLNet achieves state-of-the-art results on various NLP tasks, making it a highly effective language model.



```
0%                                             5/2196 [00:01<11:09, 3.27it/s]
Epoch 5 of 5: 100%                             5/5 [03:07<00:00, 36.66s/it]
Epochs 0/5. Running Loss: 0.0408: 100%         138/138 [00:36<00:00, 2.60it/s]
Epochs 1/5. Running Loss: 0.0187: 100%         138/138 [00:30<00:00, 5.31it/s]
Epochs 2/5. Running Loss: 0.6789: 100%         138/138 [00:30<00:00, 5.30it/s]
Epochs 3/5. Running Loss: 0.0003: 100%         138/138 [00:31<00:00, 5.06it/s]
Epochs 4/5. Running Loss: 0.0001: 100%         138/138 [00:30<00:00, 4.96it/s]
(690, 0.16990078790248303)
```

Fig 10 : Training of XLNet

4. Training of ELECTRA – Instead of masking and predicting words, ELECTRA masks and replaces them with generated samples. A separate discriminator network is then trained to distinguish between the replaced tokens and the original ones. This setup allows for more efficient training and better utilization of computational resources. By adopting this approach it can seen in Figure 11, ELECTRA achieves competitive performance with significantly fewer pre-training steps, making it a highly efficient and effective language model.

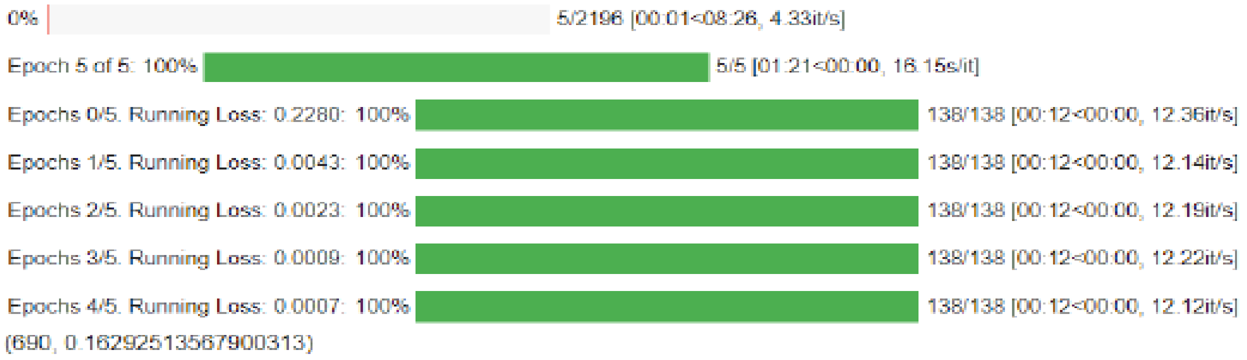


```
0%                                             5/2196 [00:00<04:08, 8.83it/s]
Epoch 5 of 5: 100%                             5/5 [02:06<00:00, 25.33s/it]
Epochs 0/5. Running Loss: 0.0240: 100%         138/138 [00:20<00:00, 7.60it/s]
Epochs 1/5. Running Loss: 0.0109: 100%         138/138 [00:20<00:00, 7.42it/s]
Epochs 2/5. Running Loss: 0.0057: 100%         138/138 [00:20<00:00, 7.71it/s]
Epochs 3/5. Running Loss: 0.0018: 100%         138/138 [00:20<00:00, 7.69it/s]
Epochs 4/5. Running Loss: 0.0022: 100%         138/138 [00:20<00:00, 7.32it/s]
(690, 0.13102528677470443)
```

Fig 11: Training of ELECTRA

5. Training of DistilRoBERTa – DistilRoBERTa utilizes knowledge distillation, where the teacher model (RoBERTa) guides the training of the student model (DistilRoBERTa) to learn the same representation as the teacher. The training process shown in Figure 12 involves minimizing the difference between the output probabilities of the teacher and student models.

DOI: https://doi.org/10.17762/ijritcc.v11i7.7938
Article Received*: 05 May 2023* ***Revised****: 26 June 2023* ***Accepted****: 15 July 2023*

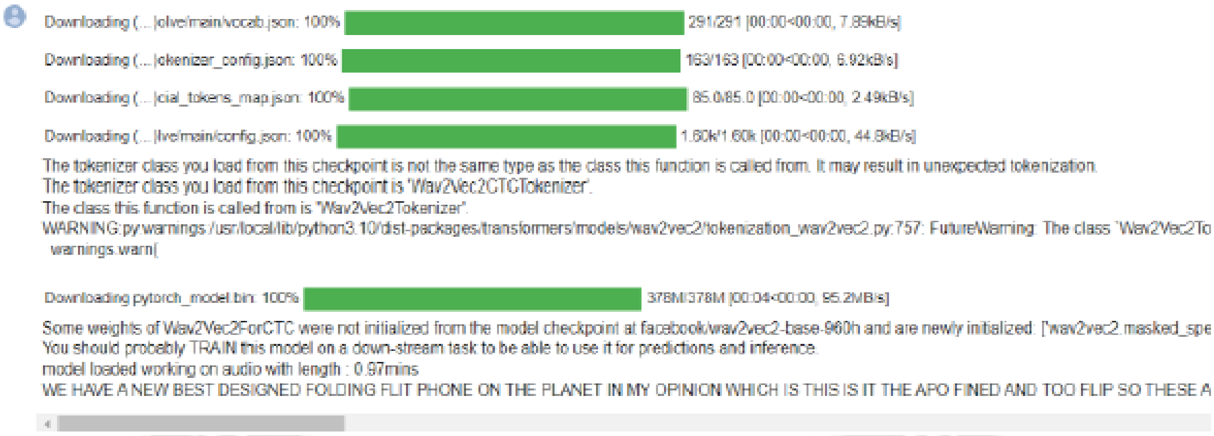


```
0%                                                   5/2196 [00:01<08:26, 4.33it/s]
Epoch 5 of 5: 100%                                   5/5 [01:21<00:00, 16.15s/it]
Epochs 0/5. Running Loss: 0.2280: 100%               138/138 [00:12<00:00, 12.36it/s]
Epochs 1/5. Running Loss: 0.0043: 100%               138/138 [00:12<00:00, 12.14it/s]
Epochs 2/5. Running Loss: 0.0023: 100%               138/138 [00:12<00:00, 12.19it/s]
Epochs 3/5. Running Loss: 0.0009: 100%               138/138 [00:12<00:00, 12.22it/s]
Epochs 4/5. Running Loss: 0.0007: 100%               138/138 [00:12<00:00, 12.12it/s]
(690, 0.16292513567900313)
```

Fig 12: Training of DistilRoBERTa

B. Transcript Generation Output

Transcribing spoken words from an audio recording into written text is known as transcription generation from audio. Automatic speech recognition (ASR) systems are frequently used for this activity since they are proficient at recognising and transcribing spoken information using complex algorithms and models. The procedure involves processes like feature extraction, acoustic modelling, language modelling, and decoding, which taken together allow the ASR system to produce precise transcriptions from audio recordings. Applications for the creation of transcriptions from audio may be found in many fields, such as transcription services, voice assistants, video captioning, and more. The use of a pretrained model for transcript creation from audio is depicted below in Figure 13.



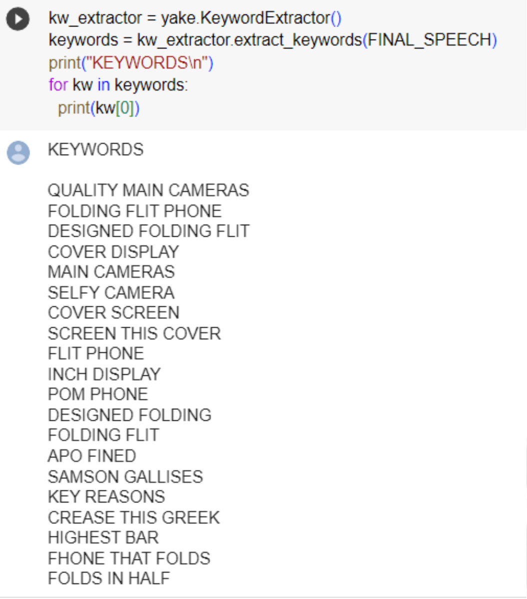
```
Downloading (…)olve/main/vocab.json: 100%            291/291 [00:00<00:00, 7.89kB/s]
Downloading (…)okenizer_config.json: 100%            163/163 [00:00<00:00, 6.92kB/s]
Downloading (…)cial_tokens_map.json: 100%            85.0/85.0 [00:00<00:00, 2.49kB/s]
Downloading (…)lve/main/config.json: 100%            1.60k/1.60k [00:00<00:00, 44.8kB/s]
The tokenizer class you load from this checkpoint is not the same type as the class this function is called from. It may result in unexpected tokenization.
The tokenizer class you load from this checkpoint is 'Wav2Vec2CTCTokenizer'.
The class this function is called from is 'Wav2Vec2Tokenizer'.
WARNING:py.warnings:/usr/local/lib/python3.10/dist-packages/transformers/models/wav2vec2/tokenization_wav2vec2.py:757: FutureWarning: The class `Wav2Vec2To
  warnings.warn(

Downloading pytorch_model.bin: 100%                  378M/378M [00:04<00:00, 95.2MB/s]
Some weights of Wav2Vec2ForCTC were not initialized from the model checkpoint at facebook/wav2vec2-base-960h and are newly initialized: ['wav2vec2.masked_spe
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
model loaded working on audio with length : 0.97mins
WE HAVE A NEW BEST DESIGNED FOLDING FLIT PHONE ON THE PLANET IN MY OPINION WHICH IS THIS IS IT THE APO FINED AND TOO FLIP SO THESE A
```

Fig 13: Transcript generation output

C. Key-Word Extraction from Transcript

Keyword extraction from a transcript involves the process of identifying and extracting important keywords or key phrases that capture the main themes, concepts, and relevant information within the text. By utilizing techniques such as TF-IDF, N-gram analysis, or machine learning-based approaches, the aim is to automatically identify the most significant terms that contribute to the overall meaning and context of the transcript. Effective keyword extraction enables researchers and practitioners to gain quick insights into the content, facilitate information retrieval, and support various applications such as summarization, categorization, and topic modeling. Below in Figure 12 is showing the implementation.

DOI: https://doi.org/10.17762/ijritcc.v11i7.7938
***Article Received**: 05 May 2023 **Revised**: 26 June 2023 **Accepted**: 15 July 2023*



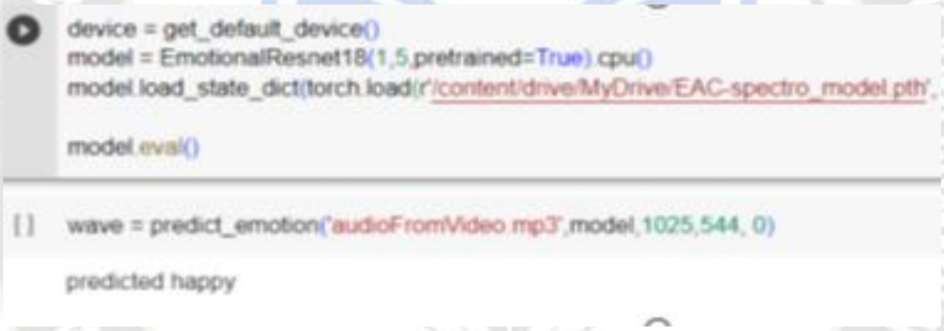
```
kw_extractor = yake.KeywordExtractor()
keywords = kw_extractor.extract_keywords(FINAL_SPEECH)
print("KEYWORDS\n")
for kw in keywords:
  print(kw[0])
```

```
KEYWORDS

QUALITY MAIN CAMERAS
FOLDING FLIT PHONE
DESIGNED FOLDING FLIT
COVER DISPLAY
MAIN CAMERAS
SELFY CAMERA
COVER SCREEN
SCREEN THIS COVER
FLIT PHONE
INCH DISPLAY
POM PHONE
DESIGNED FOLDING
FOLDING FLIT
APO FINED
SAMSON GALLISES
KEY REASONS
CREASE THIS GREEK
HIGHEST BAR
FHONE THAT FOLDS
FOLDS IN HALF
```

Fig 14: Key Word Extraction from Transcript

D. Prediction of Sentiment from Audio

Prediction of sentiment from audio is as shown in figure 15.



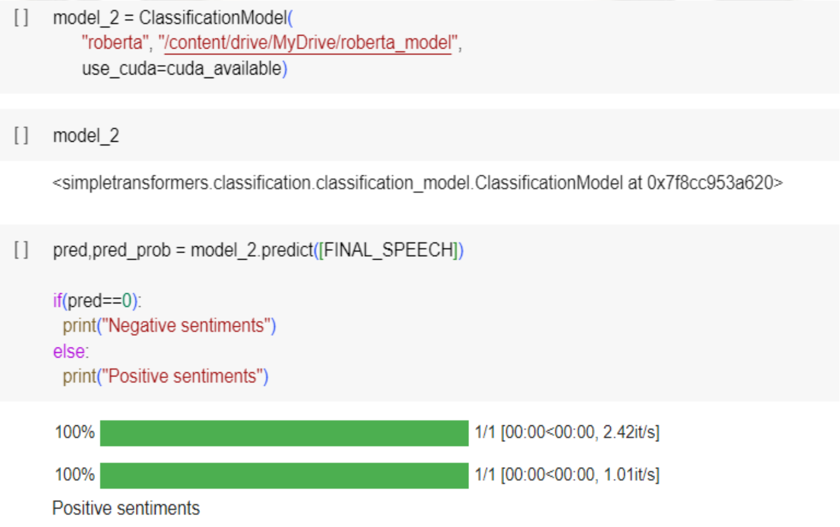
```
device = get_default_device()
model = EmotionalResnet18(1,5,pretrained=True).cpu()
model.load_state_dict(torch.load(r'/content/drive/MyDrive/EAC-spectro_model.pth',

model.eval()
```

```
[ ]  wave = predict_emotion('audioFromVideo.mp3',model,1025,544, 0)

predicted happy
```

Fig 15: Prediction of Sentiment from Audio

E. Prediction of Sentiment from Transcript

Prediction of sentiment from Transcript is as shown in figure 16.



```
[ ]  model_2 = ClassificationModel(
         "roberta", "/content/drive/MyDrive/roberta_model",
         use_cuda=cuda_available)
```

```
[ ]  model_2

<simpletransformers.classification.classification_model.ClassificationModel at 0x7f8cc953a620>
```

```
[ ]  pred,pred_prob = model_2.predict([FINAL_SPEECH])

     if(pred==0):
       print("Negative sentiments")
     else:
       print("Positive sentiments")
```

```
100%  1/1 [00:00<00:00, 2.42it/s]

100%  1/1 [00:00<00:00, 1.01it/s]
Positive sentiments
```

Fig 14. Prediction of Sentiment from Transcript

VII. Conclusion

The primary objective of this work is to provide customers with accurate and reliable ratings and reviews by leveraging deep learning models and different use cases, such as incorporating text keywords. By recognizing the importance of providing customers with a satisfying experience when browsing and purchasing products online; the aim is to enhance the overall customer experience is achieved. For audio modeling, proposed work utilized the well-known Ravdess dataset, which contains recordings of various emotional expressions, including neutral, happy, sad, angry, fearful, disgust, and surprise. The Convolutional Neural Network (CNN) based models used for various operations including to analyze spectrograms and keywords and recognize speech patterns to predict the user's sentiment accurately. The developed robust and scalable model accurately detect emotions and provide appropriate ratings and reviews to the customers. The insights and results derived from this can be leveraged to enhance the overall customer experience, thus leading to higher customer satisfaction and loyalty.

To enhance the project further, several improvements can be considered: Augmentation Techniques to augment the spectrograms and Model Accuracy Enhancement by exploring different network architectures, fine-tuning hyper-parameters, and incorporating ensemble methods to combine multiple models' predictions

References

[1] Shiqing Zhang , Shiliang Zhang , Member, IEEE, Tiejun Huang , Senior Member, IEEE, and Wen Gao, Fellow, Speech Emotion Recognition Using Deep Convolutional Neural Net- work and Discriminant Temporal Pyramid Matching IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 20, NO. 6, JUNE 2018

[2] M. E. Seknedy and S. Fawzi, "Speech Emotion Recognition System for Human Interaction Applications," 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), 2021, pp. 361-368, doi: 10.1109/ICICIS52592.2021.9694246

[3] JORGE OLIVEIRA AND ISABEL PRAÇA On the Usage of Pre-Trained Speech Recog- nition Deep Layers to Detect Emotions Received December 22, 2020, accepted January 1, 2021, date of publication January 12, 2021, date of current version January 19, 2021. Digi- tal Object Identifier 10.1109/ACCESS.2021.3051083

[4] Sabrine Dhaouadi,Hedi Abdelkrim,Slim Ben Saoud Speech Emotion Recognition: Models Implementation & Evaluation

[5] M. Gokilavani, H. Katakam, S. A. Basheer and P. Srinivas, "Ravdness, Crema-D, Tess Based Algorithm for Emotion Recognition Using Speech," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 1625-1631, doi: 10.1109/ICSSIT53264.2022.9716313

[6] X. Liu, Y. Mou, Y. Ma, C. Liu and Z. Dai, "Speech Emotion Detection Using Sliding Window Feature Extraction and ANN," 2020 IEEE 5th International Conference on Signaland Image Processing (ICSIP), 2020, pp. 746-750, doi: 10.1109/ICSIP49896.2020.9339340

[7] SUDARSANA REDDY KADIRI , (Member, IEEE), AND PAAVO ALKU , (Fellow, IEEE) Received March 5, 2020, Excitation Features of Speech for Speaker Specific Emotion Detection accepted March 20, 2020, date of publication March 24, 2020, date of cur- rent version April 9, 2020. Digital Object Identifier 10.1109/ACCESS.2020.2982954

[8] REZA LOTFIAN , (Student Member, IEEE), AND CARLOS BUSSO , (Senior Member, IEEE) , Lexical Dependent Emotion Detection Using Synthetic Speech ceived January 14, 2019, accepted January 29, 2019, date of publication February 8, 2019, date of current version March 1, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2898353

[9] Tyagi, R. ., K. Shastri, R. ., M., K. ., Ramkumar Prabhu, M. ., Laavanya, M. ., & C. Pawar, U. . (2023). Undecimated Wavelet Transform Technique for the Security Improvement In the Medical Images for the Atatck Prevention. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 211–217. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2563

[10] S. Prasomphan, "Detecting human emotion via speech recognition by using speech spectrogram," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015, pp. 1-10, doi: 10.1109/DSAA.2015.7344793.

[11] Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORK CONSIDERING VERBAL AND NONVERBAL SPEECH SOUNDS

[12] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America, 8, 185–190

[13] R. Cowie et al., “Emotion recognition in human-computer interaction,” IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32–80, Jan. 2001.

[14] Andrew Hernandez, Stephen Wright, Yosef Ben-David, Rodrigo Costa, David Botha. Optimizing Resource Allocation using Machine Learning in Decision Science. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/195

[15] S. Ramakrishnan and I. M. El Emary, “Speech emotion recognition approaches in human computer interaction,” Telecommun. Syst., vol. 52, no. 3, pp. 1467–1478, 2013.

[16] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: features, classification schemes, and databases,” Pattern Recogn., vol. 44, no. 3, pp. 572–587, 2011.

[17] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011,” Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177, 2015.

[18] F. Eyben et al., “The Geneva minimalistic acoustic parameter

set (GeMAPS) for voice research and affective computing," IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016

[19] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278– 2324, Nov. 1998

[22] Mr. Ather Parvez Abdul Khalil. (2012). Healthcare System through Wireless Body Area Networks (WBAN) using Telosb Motes. International Journal of New Practices in Management and Engineering, 1(02), 01 - 07. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/4

[23] R. Govindwar et al., "Blockchain Powered Skill Verification System," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-8, doi: 10.1109/ICONAT57137.2023.10080848

[24] Khetani, V. ., Gandhi, Y. ., Bhattacharya, S. ., Ajani, S. N. ., & Limkar, S. (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 253–262.