Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Multimodal Speech Sentimental Analysis

[1]Rohan Katyal, [2]Sanskriti Agrahari, [3]Shagun Chauhan, [4]Md. Shahid,

Department of CSE, MIET, Meerut,

***Abstract* —** In today's world, communication is not just limited to text. Communication today can be in the form of various modes, such as speech, visual, and textual cues. With the help of these multiple modalities, sentiment analysis becomes more easy and accurate. This paper introduces Multimodal Speech Sentiment Analysis (MSSA), which integrates Convolutional Neural Network (CNN) and Bidirectional Encoder Representations from Transformers (BERT) algorithms for comprehensive analysis of the sentiments that are expressed through multiple modalities, such as acoustic features from speech. This paper sets a new standard for sentiment analysis by promoting speech data along with conventional textual analysis. We are not only focusing on multimodal approaches so that we can interpret the sentiments accurately but also seizing the complications of human expressions in the real world.

***Keywords* —** Bidirectional Encoder Representations from Transformers (BERT); Multimodal Sentiment Analysis (MSA); Speech Representation Learning (SRL); Deep Learning (DL); Convolutional Neural Network (CNN), Ensemble Learning.

1. Introduction

The primary means of communication is speech, which holds a wealth of sentimental information. It is very crucial for various applications to understand the emotions that are expressed behind the spoken words. With technological advancements, the blending of emotions and computational intelligence has become a highlight in various domains. A diversity of human expressions— voices, gestures, and facial cues—interact with the intelligent systems. In conventional techniques, we were primarily focusing on the written text, but now that we incorporate speech data into analysis models, it improves both accuracy and depth.

Multimodal speech sentiment analysis's (MSSA) primary goal is to interpret human emotions by concurrently examining auditory and visual cues. It has the ability to grasp non-oral hints such as tone, pitch, and volume. These hints are really helpful as they provide precious perceptions into the speaker's mental state, thereby upgrading sentiment analysis precision. It is really important to understand the emotions behind the spoken language, as it can be used for numerous applications such as customer feedback analysis, market research, and mental health monitoring. This can be achieved by integrating the advanced algorithms named Convolutional Neural Networks (CNN) and Bidirectional Encoder Representations from Transformers (BERT) models.

This research paper begins a thorough investigation, delving into the intricate varieties of multimodal sentiment analysis across these various domains. This endeavor seeks to unearth the depth and breadth of sentiments within human- computer interactions, entertainment mediums, and the nuanced realm of political discourse through a synthesis of empirical studies, theoretical frameworks, and practical applications. MSSA (Multimodal Speech Sentiment Analysis) is a sophisticated approach to understanding and interpreting human expression that examines sentiments conveyed via multiple modes of communication. This innovative approach integrates several data sources, including text, audio, and visual clues, to improve comprehension of the feelings and emotions people convey. MSSA's primary goal is to overcome the limitations of traditional sentiment analysis, which primarily focuses on textual data.

1.1 Background

A. Speech Recognition

Speech recognition is a technology that converts the language spoken by humans into written text. Speech Recognition is also known as Automatic Speech Recognition (ASR), whose fundamental aim is to enable computers or audio-controlled devices to understand human speech. The process of speech recognition includes complex signal processing, pattern recognition, and language modeling in order to precisely translate the spoken words. Speech recognition is generally confused with voice recognition. Speech recognition focuses on the translation of speech from oral format to text format, while voice recognition looks to distinguish a singular client's voice. Speech is also utilized for speaking with shrewdhome machines [9].

The various components of Speech Recognition are: (a) Acoustic Signal Processing: In this process, sound waves in the air are captured by the microphones or other transducers, thereby converting them into electrical signals.

(b) Pre-processing and Feature Extraction: In order to remove the noise and enhance the signal quality, the raw acoustic signals are preprocessed. After preprocessing, feature extraction is applied to the signal to capture features such as pitch, intensity, and spectral content (Common features: Mel-Frequency Cepstral Coefficients (MFCCs) or filter banks). (c) Acoustic modeling: A statistical relationship is developed between extracted features from acoustic signals and phonemes. Conventionally, Hidden Markov Models (HMMs) were used for this purpose, but recently, Deep Neural Networks (DNNs) have proved to be more effective in terms of performance. (d) Language Modeling: The likelihood of a particular word occurring in a given context is calculated. N-gram models, Recurrent Neural Networks (RNNs), and Transformer models are used for language modeling. (e) Decoding: The acoustic features are aligned with the most likely sequence of linguistic units. (f) Post-Processing: It involves error-detection, content-based adjustments, etc. Inthis paper, we are going to utilize Sphinx to have text from the speech modality.

B. Sentiment Analysis

Sentiment Analysis is a Natural Language Processing (NLP) technique that recognizes the emotions and sentiments behind a text. The sentimental analysis helps us identify the emotions, beliefs, attitude, goals, ideas, and opinions that are expressed in a text. Sentiment Analysis is also termed opinion mining. The most fundamental type of sentiment analysis is binary classification, i.e., either positive or negative [2]. Further, the sentiments can also be categorized as positive (favorable/satisfaction), negative(unfavorable/dissatisfaction), or neutral (lack of a clear emotional tone). Strategies based on knowledge, statistical, and hybrid approaches are the three methodologies that make up sentiment analysis. Text is distinguished by strategies based upon knowledge using unambiguous influential classifications such as "sprightly" (indicating positive sentiment), "hopeless" (indicating a negative sentiment), "froze" (potentially indicating a negative or neutral sentiment), and "unremarkable" (indicates a neutral or negative sentiment). There is also plausible "fondness" for certain sentiments. Statistical Methods make use of machine learning techniques so that patterns and relationships between textual features and sentiment labels are learned automatically. Semantic space models, deep learning, latent semantic analysis (LSA), word embedding models, and support vector machines (SVM) are examples of machine learning techniques (MLT) that are utilized in statistical methodologies. Hybrid approaches make use of both machine learning techniques and elements from knowledgerepresentation.

A lot of research has been done on sentiment analysis by the researchers, who mainly focus on "Textual Sentiment Analysis." However, "Audio Sentiment Analysis" is in its early phases and has begun to intrigue the technology world.. Audio Sentiment Analysis is a subsection of Natural Language Processing (NLP) that pays attention to extracting and understanding sentiments, attitudes, and opinions that are conveyed through spoken language. By using various machine learning algorithms and signal processing techniques, technologists focus on establishing booming models that can precisely recognize sentiments from audible data [3] [4]. Multimodal speech sentiment analysis is a swiftly advancing field that focuses on providing a detailed and nuanced

understanding of human sentiments that are communicated through speech by merging various modalities like looks, motions, and acoustic prompts [1].

C. Deep Learning

The modern neural network is made up of small computing units. Every unit takes a vector of values as input and gives a single value as an outcome. It is because of this framework that neural networks learn complicated patterns and representations from data. The applications of modern neural networks are frequently referred to as deep learning. The word "deep" implies that the network has numerous layers, thereby giving it the potential to learn complex hierarchical representations of data [2].

Well-known deep learning techniques, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have greatly impacted fields like computer vision and natural language processing. Essentially, convolutional neural networks (CNNs) leverage the spatial correlations between pixels to recognize images with grid-like structures. In Convolutional Neural Networks (CNNs), filters or kernels are applied to images to fetch suitable attributes [7] [8]. Recurrent Neural Networks (RNNs) contain a cycle within their own network connections, thereby capturing and processing sequential information [2].

D. Machine Learning

Machine learning is a part of artificial intelligence that allows computing systems to acquire knowledge from data and enhance their execution overtime without being explicitly programmed. Machine learning spins around the evolution of algorithms that make computing systems understand patterns and their data driven results. These algorithms are instructed using huge quantity of data which can be labeled (Supervised learning), unlabeled (Unsupervised learning) or require response loops (Reinforcement learning). Supervised learning demands training models on labeled data, where all feeds are coupled with a corresponding output. Here, the model acquires the knowledge of plotting inputs to outputs, allowing it to make forecast on unseen data. Unsupervised learning deals with unlabeled data, trying to find the unseen structures within the data without any external directions. Ordinary tasks in unsupervised learning are clustering and dimensionality reduction. Reinforcement learning emphasizes on instructing agents to link with the surroundings to reach a certain aim. Through trial run, the agents acquires knowledge that which steps give most commending results, refining its judging process over time. Various algorithms used in machine learning are: linear regression, logistic regression, decision trees, random forest, support vector machines (SVM), K nearest neighbors (KNN), neural networks, naive bayes, clustering algorithms (example K-means, hierarchical clustering), principle component analysis (PCA). Applications of machine learning are: healthcare, finance, marketing, robotics etc. [9].

E. Ensemble Learning

Ensemble learning is about integrating predictions from various models to generate a concluding prediction. This approach applies the collective intelligence of varied models to get fine execution than any independent model could alone. Numerous ensemble methods are as follows: bagging (example - random forest), boosting sequentially (example - gradient boosting machines), stacking often called meta-learner or blender, voting (it takes majority votes for classification purpose and averaging for regression purpose). In our project, we have used stacking and voting methods to combine CNN and BERT models. It is due to ensemble learning that risk of over fitting is minimized and there is an improvement in generalization. Ensemble learning is generally used in classification, anomaly detection, model combination etc.

1.2 Supported technologies and algorithms

We will utilize Jupyter Notebook and Anaconda Navigator to implement Multimodal Speech Sentiment Analysis (MSSA). Anaconda Navigator's graphical interface makes it simple to use. For non-identical projects, we can construct well-defined isolated settings to ensure that each project operates independently to avoid overlapping. Jupyter Notebook is an open-source web tool that allows users to create and share formulas, documents with live code, pictures, and explanation text. It is a versatile platform that frequently provides

interactive settings for machine learning efforts, data analysis projects, and scientific study. It is a powerful tool that is widely used in educational settings and research sectors [10].

We will use the Multimodal Speech Sentiment Analysis (MSSA) technique, which is backed by two primary algorithms: (a) Convolutional Neural Network (CNN) and (b) Bidirectional Encoder Representations from Transformer (BERT). Image processing and computer vision fields frequently use Convolutional Neural Networks (CNNs), which are advanced deep learning techniques.Convolutional Neural Networks (CNNs) were inspired by the visual systems of anthropoids. Visual systems, also known as Convolutional Neural Networks (CNNs), are responsible for extracting meaningful attributes from complex visual inputs. When it comes to analyzing sentiments, opinions, and beliefs in human speech, Convolutional Neural Networks (CNNs) have proven to be significantly more powerful than standard neural network approaches [11]. Bidirectional Encoder Representations from Transformers (BERT) is renowned for properly understanding the context, which has earned it recognition in the field of emotive analysis. To perform sentiment analysis on multimodal voice data, the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model is fine-tuned. The model's strength is in understanding and analyzing the complex relationships between words and their context, allowing it to detect minute differences in sentiment with high accuracy [12].

2. Proposed Work Plan

2.1 General Architecture/ Flowchart/ DFD

Multimodal speech sentiment analysis is a branch of natural language processing (NLP).We will use multimode (i.e., speech with text) instead of just speech in order to gain a thorough understanding of the sentiments being conveyed. With the growing usage of social media platforms, analyzing and interpreting the sentiments and emotions behind spoken words has become more complex for a diversity of applications such as healthcare, financial trading, and market research [13]. As illustrated in Figure 1, multimodal speech sentiment analysis consists primarily of two processes: multimodalspeech recognition and sentiment analysis.

In multimodal speech recognition, speech is first broken down into syllables. The syllables are thenconverted to text for sentiment analysis. After generating the text, it will be processed and analyzed to provide a result that can be either positive or negative (binary classification). The sentiments analyzed by the machine are included in the results. Multimodal speech sentiment analysis will unite both audio and linguistic data in order to acquire a high- quality understanding of the emotional expression behind the spoken language. To fulfill this, researchers integrated convolutional neural networks (CNN) and bidirectional encoder representations from transformer (BERT) models.

Convolutional Neural Network (CNN) models are excellent at pulling out attributes from images and videos, making them worthy of analyzing visual hints in multimodal sentiment analysis. On the other hand, Bidirectional Encoder Representations from Transformer (BERT) models are capable of processing textual data by seizing contextual information. By uniting the above algorithms, a multimodal speech sentiment analysis model can successfully capture both linguistic content and non- oral emotional hints from audio-visual inputs [14].

Convolutional neural networks (CNN) have turned out to be an influential tool for multimodal speechsentiment analysis (MSSA). Convolutional neural networks (CNNs) are intended to imitate the visual cortex of the anthropoid brain, allowing them to bring out the important attributes from raw input data. Audio signals are divided into small segments called frames,which are analyzed by convolutional neural networks (CNNs). There are multiple convolutional layers through which each and every frame is passed. These convolutional layers apply filters to recognize several patterns within the audio data. Different acoustic properties, such as pitch, timbre, and intensity, are captured by the filters. The outputs obtained after filtration from the convolutional layers are then supplied to fully connected layers (high-level representations of emotions present in speech). Now appropriate activation functions (such as the softmax function) will be utilized to process and classify these representations.

Bidirectional Encoder Representations from Transformer (BERT) is a remarkable model that was developed by Google and is used for various Natural Language Processing (NLP) problems, including embedding. A high-powered approach is offered by Bidirectional Encoder Representations from Transformer (BERT) so that the sentiments behind the spoken words can be understood more accurately. Transformer-based architectures are used by Bidirectional Encoder Representations from Transformer (BERT) in order to get and understand the contextual connection among words in a data file (both left and right context are considered) [15]. Bidirectional encoder representations from the Transformer (BERT) model, when pre-trained on a large unlabeled dataset and fine-tuned for certain applications (such as sentiment analysis, language translation, and text categorization), can produce futuristic results. For speech sentiment analysis, acoustic attributes fetched from words spoken by humans and textual data (capturing emotions and nuances) are processed by Bidirectional Encoder Representations from Transformer (BERT). The encoder network learns to fetch valuable representations from auditory data while concurrently incorporating lingual context.

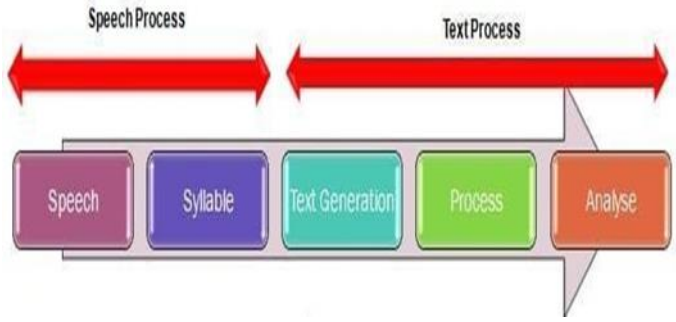

Figure-1: Speech and Text SentimentalAnalysis

Figure 2 represents how, by integrating the above two high- powered algorithms, researchers have gained outstanding performance in sentiment classification. Future research must pay heed to problems such as having multimodal data that is not well labeled and upgrade the interpretability of hierarchical learning [16].

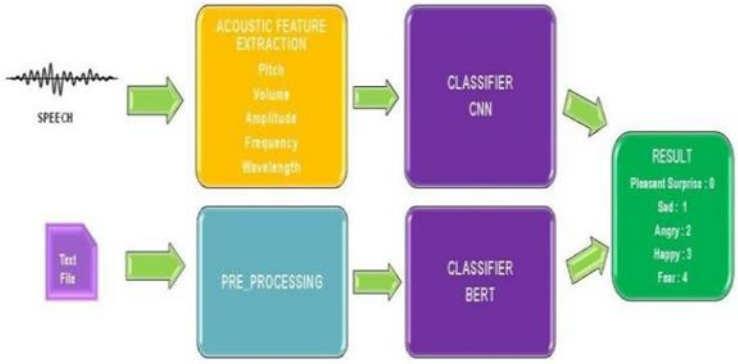

Figure-2: Multimodal Speech Sentimental Analysisworkflow

2.2 Description of various modules

A. Transformers

Transformers are good at managing complex language structures, understanding interdependencies, and providing upgraded representations of NLP applications. They fetch nuances from expressions and irony. They are the basis for NLP transfer learning. It provides impressive sentiment analysis features to judge emotional polarity. Various pre- trained transformers are as follows: BERT (Bi-Directional Encoder Representation from Transformers), GPT-2 (Generative Pre-Trained Transformers), or RoBERTa (Robustly Optimized BERT Approach). The textual data is pre-processed with the help of tokenization (the division of text into small units such as words or sub- words). The tokens are then fed to the transformer, which results in contextual

representation for each and every token.

B. Tensorflow

Tensorflow is an open-source machine learning software provided by Google. Tensorflow makes it trouble-free to create advanced deep learning models by providing a flexible tech environment. Tensor Flow's key structure is a data flow graph, where nodes depict mathematical operations and edges depict tensors— multidimensional data arrays. It enables parallelism, thereby allowing Tensorflow to use CPUs or GPUs. This different design enables parallelism, allowing TensorFlow to use resources such as CPUs or GPUs. Its interface supports, Python, Java, C++ and other different programming languages[18].

C. PyTorch

It is a machine learning library that has recently become well known because of its efficiency. It helps in the successful development of deep learning models such as Multimodal Speech Sentiment Analysis (MSSA) [19]. It is available as open-source and has the capability to handle complicated data representations. PyTorch is effective where we need to combine multiple modalities such as textual content, acoustic features, and facial expressions to enhance the accuracy of sentiment analysis.

2.3 Algorithm of main component

As mentioned earlier, our project is composed of two main components:

1. **Multimodal Speech Recognition and**
2. **Sentiment Analysis**

2.3.1 Multimodal Speech Recognition

The algorithm used for implementing multimodal speech recognition is a convolutional neural network (CNN), whereas the algorithm used for implementing sentiment analysis is bidirectional encoder representations from transformers (BERT).

Convolutional Neural Network (CNN) captures dependencies such as time and spatial arrangement of pixels, thereby capitalizing on temporal and spatial connections existing in both audio and visual streams. Bidirectional Encoder Representations from Transformers (BERT) was introduced in 2018 by Google. It uses transformer-based architectures. It is pre-trained and fine- tuned for particular tasks, such as sentiment analysis, where it takes into account contextual connections and semantic meaning in the given text. By utilizing several modalities, multimodal speech recognition seeks to improve the resilience and accuracy of speech recognition algorithms. It is a subfield of artificial intelligence (AI). There are several obstacles that are caused by noisy environments, accents, or speaker variations. Multimodal speech recognition systems solve these obstacles by merging acoustic signals with visual details like facial expressions or lip movement. For multimodal speech recognition, convolutional neuralnetworks (CNNs) are used because of their deep learning capability [20].

We are implementing speech recognition with the helpof convolutional neural networks (CNNs) because they have the capability to fetch local dependencies (relationships among contiguous audio frames) in data and pull out high- level attributes from audio signals. In multimodal speech recognition, the audio and visual datasets are processed concurrently. Hierarchical learning is made possible in convolutional neural networks (CNNs) by employing multiple convolutional layers. Each and every convolutional layer fetches different complex patterns and variations in speech signals. Convolutional neural networks (CNNs) possess the ability to handle spatial and temporal characteristics among audio frames, thereby improving the accuracy of speech recognition [21].

A convolutional neural network (CNN) is also called a "local network" because, at a particular point in time, it primarily focuses on local parts of the input made by the user. Each unit is calculated in the precise "window" location that is influenced by the region that the user is now seeing. The main layers and responsibilities of the forward feed structure of a convolutional neural network (CNN) are:

1. Convolutional Layer: It extracts attributes frominput data.
2. Sub-sampling Layers: They lowerthe dimensionality of the input data.
3. Layer of aggregation (pooling): Here, poolingfurther reduces the dimensionality of the data.
4. Fully-Connected Layer :It possessesthe responsibility to predict the classes.

Each convolutional layer plane is linked to one or more feature maps from the preceding layer, thus the specifics of the identified features are transferred from one layer to the next. The output of a plane (2-D matrix or feature map) is now obtained by applying an activation function. One or more feature maps are produced by a convolutional layer. In the next sub- sampling (pooling) layer, each feature map is now associated with a single plane.[22]

2.3.2 Sentiment Analysis

Here we will have textual data, and we are going to use the Bidirectional Encoder Representations from Transformers (BERT) algorithm to analyze and understand that text. It is an automated process that determines the sentiments conveyed through the text. Textual sentiment analysis is often known as opinion mining. The algorithm is responsible for analyzing the emotions behind the tone, attitude, or opinion that is conveyed through the written text. The successful interpretation of the emotions behind the text can be applied in various fields mentioned as following social media monitoring, brand monitoring and reputation management, customer feedback analysis, product and service reviews, financial news analysis, political sentiment analysis, employee feedback and engagement, healthcare patient feedback, educational

feedback analysis, tourism and hospitality, legal document analysis, and chatbot and virtual assistant training [24].

All these applications focus on the visualization of sentiment analysis based upon bidirectional encoder representations from transformers (BERT) across numerous business industries and use cases. Bidirectional Encoder Representations from Transformers (BERT) is helpful in situations where context understanding is critical to deciding sentiment polarity (the most basic is the binary classification, where polarity can either be positive or negative). Bidirectional Encoder Representations from Transformers (BERT) has the ability to understand the scope of negation words (example: not, never) depending on the polarity of the sentiment. With the help of the surrounding context, Bidirectional Encoder Representations from Transformers (BERT) can precisely decide whether the statement express a positive polarity or negative polarity.

The conventional language models were used to process and analyze the text in one direction (unidirectional, from right to left or left to right). The keyword “bidirectional” in Bidirectional Encoder Representations from Transformers (BERT) indicates the capability of the algorithm to consider words in a sentence from both sides, i.e., left to right as well as right to left. Bidirectional Encoder Representations from Transformers (BERT) is based upon the encoder-only architecture of the transformer model. Itis pre-trained on a huge dataset using unsupervised learning, where it tries to guess in advance the words that are missing in a sentence. Bidirectional Encoder Representations from Transformers (BERT) not only impacts sentiment analysis but also impacts named entity recognition, question answering, and language understanding tasks of natural language processing (NLP).

After pre-training, fine-tuning takes place. The aim of the algorithm is its ability to take into accounts the complicated connections and contextual surroundings of words in order to find out subtle nuances in sentiment with outstanding accuracy. Bidirectional Encoder Representations from Transformers (BERT) has the ability to understand the scope of negation words (example: not, never) depending on the polarity of the sentiment. The performance of Bidirectional Encoder Representations from Transformers (BERT) is calculated by comparing its accuracy, precision, recall, F1 score, and other applicable criterions against annotated datasets [25].

3. Experimental ResultAnalysis

3.1Description of data set used.

To implement and run our entire model, we use a total of four datasets. The Toronto Emotional Speech Set (TESS), Audio Speech Sentiment, and Emotion Detection from Text and LibriSpeech are the four datasets. The TESS dataset contains a set of 200 target words spoken in the carrier phrase "Say the word _" by two actresses (aged 26) such a way that each of the two female actors and their emotions has its own folder. And within that, all 200 target word audiofiles can be found. The audio files are present in WAV forms. The goal of this dataset is to test the CNN model.

Audio Speech Sentiment Data consists of 721 different audio files containing various emotions that can be classified into three categories: positive, negative, and neutral. This dataset will not be used to train any models. It is used to provide input data to our models, so it is used during the testing phase of the model. Emotion Detection from Text Data is essentially a collection of tweets that have been annotated with the emotions associated with them. There are three columns in this table: tweet_id, sentiment, and content. The raw tweet is in "content." "Sentiment" refers to the emotion behind the tweet. We have 13 different emotions and forty thousand records in our data. It is only used in the BERT model for sentiment analysis. LibriSpeech is a corpus of approximately 1000 hours of 16 kHz read English speech created by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from audio books read for the LibriVox project and has been carefully segmented and aligned. It consists of speech files as well as transcript files in English. Its primary function is to convert speech to text. Converting speech to text can be accomplished in two ways: the first is to use CMU Sphinx from the SpeechRecognition library to convert speech to text for a dataset containing WAV files and a CSV filecontaining transcript data, and the second is to use Google's Speech Recognition API or any other API.

Calculate the efficiency or accuracy of thedesigned system according to the parameterused to evaluate the system.

Figures 3, demonstrate the validation accuracy and validation loss, and as per the demonstrations provided by the confusion matrix in Figure 4, it concludes that most of our sample cases were identified accurately, and we were able to get decent results from the models. Accuracy of 97.86% is achieved on the successful completion of the training of CNN model, with a loss of 6.87%. Table 1A presents more thorough results for all the models in terms of accuracy and losses of the different models. This table provides a full categorization report for all of the distinct classes. The confusion matrix demonstrates that the great majority of our samples were accurately identified.

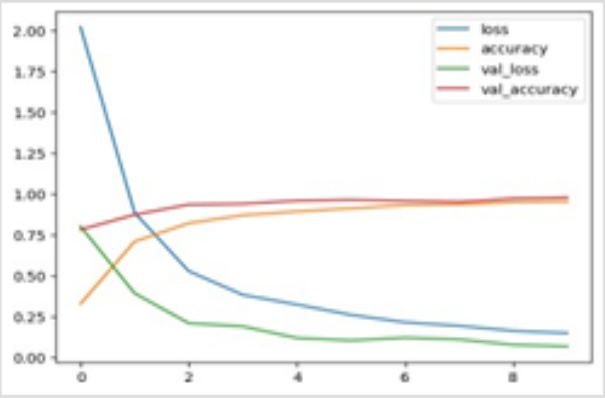

Figure-3: Validation Accuracy and loss

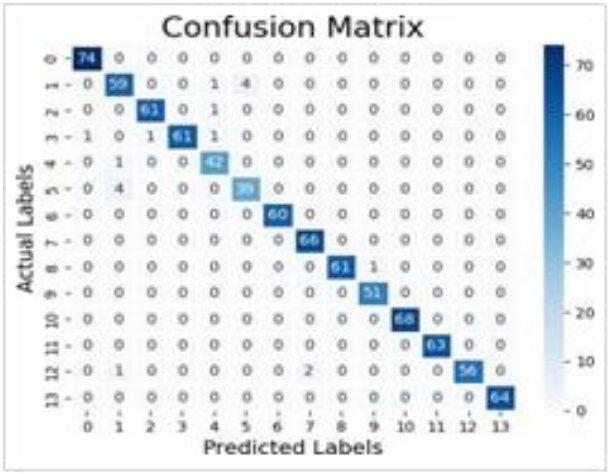

Figure-4: Confusion Matrix

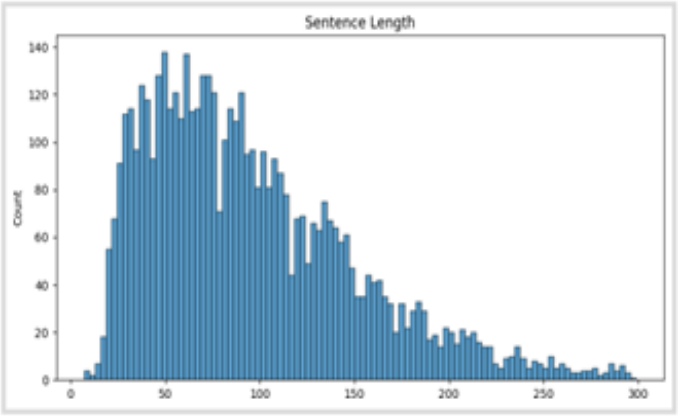

Figure-5: Sentence Length V/S Count for BERT

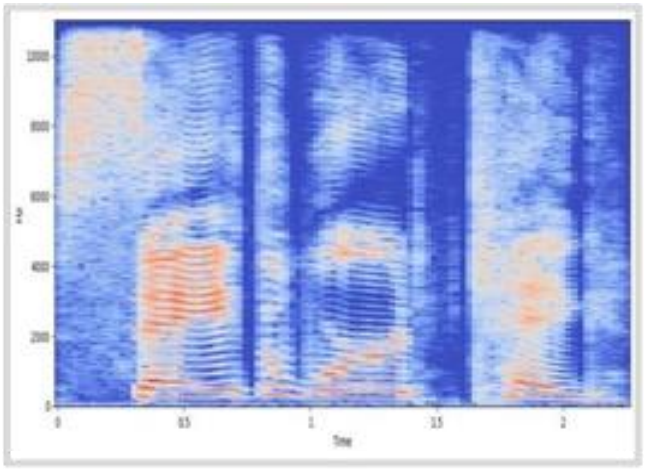

Figure-6: Specshow (spectogram show)ofan example audio

For BERT, our model has an accuracy of 91.73% with a total of ten epochs. We also obtained different F1 Score Values for different epochs. They are 89.0%, 90.6%, 90.4%, 92.2%, 92.9%, 93.4%, 93.5%, 93.5%, 93.5%, and

93.5% respectively from epoch 1 to epoch 10. The average of all these F1 Score is 92.32%. Figure 5 depicts

a graph about the sentence length in the dataset used for BERT. It is a count v/s sentence length graph in which the maximum length of the sentence is 167.

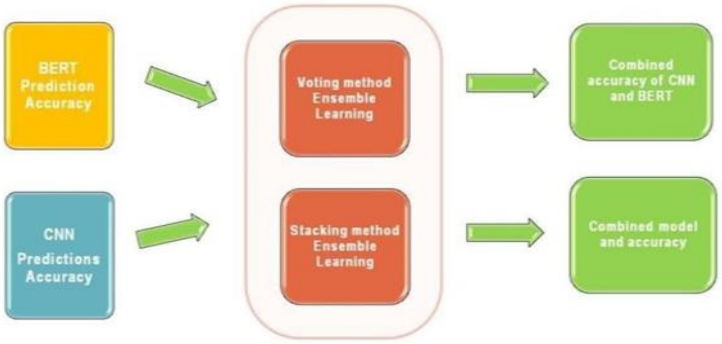

Figure-7: Combination of CNN and BERT usingEnsemble Learning

After controlling large number of trials using CNN and BERT, we come to the conclusion for the CNN model, that it's prime for anticipating emotions through audio. When it comes to the execution of the two models (i.e. CNN and BERT), we deduce that CNN outruns BERT. We utilized a speech recognizer library together with the CMU Sphnix or directly make use of a Google's API in order to transform audio data to text. We can use a transcript dataset to train this model. After that we run both CNN and BERT models together in such way that CNN is utilized for audio and BERT is utilized for text analysis. Now we examine all the models (Table 1B).

Model's Name	Accuracy	Loss
CNN	97.86%	6.87%
BERT	91.73%	8.27%
CNN+BERT	94.18%	7.71%

Table 1A: Model Performance

According to the information retrieved from above table, BERT executes badly in contrast to the audio model. Audio model executes with 91.73% accuracy, owing to its

increased vulnerability to complication with automated transcription, resulting in bad text output and a notable reduction of our model's overall execution. External factors like noise in the background (aural environment, moving vehicle, sharp noise, resonance, ding, music etc), numerous human beings talking at once, speedy speakers, and bad auditory standards are responsible for affecting the correctness of an automated transcript.

The outcome is that the multimodal model outruns the CNN model for the reason that it utilized both modes of data (auditory and text) and amalgamate all bring about a single resolution vector to get aftermaths of the sentiment classification. One edge of our technique is that it does not need the blend of varied data, and each method can utilize its most apt categorization blueprint singly.

Ensemble learning works on uniting different Predictions from our base models i.e. BERT and CNN models used in our project. The BERT prediction accuracy and CNN prediction accuracy are used in order to have much precise final prediction. Here voting ensemble learning technique is used to get the combined accuracy of CNN and BERT and the Stacking method is providing us with combined model accuracy by creating a combined model of BERT and CNN.

4. Conclusion

The research presents multimodal sentiment analysis model embodying text and audio modes as the basis for identifying and categorizing emotions. We employ combination of four distinct data sets and three different

models: BERT, CNN, and hybrid model that combines the best features of both CNN and BERT. Using the CNN model, we transform the trained BERT, by embodying auditory modality data in order to aid textual modality data, from unimodal to multimodal. We are able to get the accurate representation. Our results show that the multimodal model is one of the primes for the emotion identification, and may give rise to the developmentof multimodal models in future researches.

References:

[1] Daniel Jurafsky and James H. Martin, "Speech andLanguage Processing", Third Edition

[2] Joas Pambou, "Using AI To Detect Sentiment In AudioFiles", 2023

[3] Dr. Nisha Auti, Atharva Pujari, Anagha Desai, Shreya Patil, Sanika Kshirsagar, Rutika Rindhe, " Advanced Audio Signal Processing for Speaker Recognition and Sentiment Analysis", Volume 11 Issue V May 2023-

[4] "Introduction to Deep Learning", 2023 "The Data Science andAI Handbook – How to Start a Career inData Science",2023,

[5] "Introduction to Convolution Neural Network", 2023

[6] Vinod Sharma, "Deep Learning – Introduction to Convolutional Neural Networks"

[7] "Everything You Need to Know About AI Cyber security", 2023, "Running R in Jupyter: Unleash the Simplicity ofNotebooks", 2023.

[8] Anvarjon T, Mustaqeem and Kwon S., "Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features." Sensors (Basel). 2020, PMCID: PMC7570673.

[9] Guangyao Pang, Keda Lu, Xiaoying Zhu, Jie He, Zhiyi Mo, Zizhen Peng, Baoxing Pu,"Aspect-Level Sentimental Analysis Approach via BERT and Aspect Feature Location Model", Wireless Communications and Mobile Computing, vol.2021.

[10] Neri Van Otten, "The History Of Natural Language Processing & Potential Future Breakthroughs", 2023.

[11] Jing He, Haonan Yanga, Changfan Zhang, Hongrun Chen, Yifu Xua, "Dynamic Invariant-Specific Representation Fusion Network for Multimodal Sentiment Analysis", Computational Intelligence and Neuroscience, vol. 2022

[12] Mansi Jain, Purvit Vashishtha, Aman Satyam and Smriti Sehgal, "CNN LSTM Hybrid Approach for Sentiment Analysis" Volume 11 Issue V May 2023

[13] Yang, Q., Li, X., Ding, X. et al. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. Alz Res Therapy 14,186 (2022).

[14] Yulia Gavrilova, "Transformers in ML: What They Are and How They Work", 2023

[15] Devopedia, 2019, "TensorFlow." Version 8, November 4.Accessed 2023-11-13.

[16] Jianguo Sun, Hanqi Yin, Ye Tian, Junoeng u, Linshan Shen, Lei Chen, "Two-Level Multimodal Fucion for Sentimental Analysis in Public Security", Security and Communication Networks, vol. 2021.

[17] Francis R, Richard HA, Manda RN and Niriarijaona RH, "Voice User Interface: Literature Review, Challenges and Future Directions", 2021, Vol.9 No.10: 113.

[18] Mustaqeem and Kwon S., "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition",Sensors (Basel), 2019, PMID: 31905692.

[19] Ayad Alsobhani et al, "Speech Recognition using Convolution Deep Neural Networks", 2021, J. Phys.:Conf. Ser. 1973 012166

[20] Ilias Papastratis, "Speech Recognition: a review of thedifferent deep learning approaches", 2021.

[21] Medium, "Sentiment Analysis Using BERT",2022.

[22] Cach N. Dang, Maria N. Moreno-Garcia, Fernando De la Prieta, "Hybrid Deep Learning Models for Sentimental Analysis", Complexity, vol. 2021.