International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I1P226
Volume 69 Issue 1, 172-178, January 2021

A Speech-based Sentiment Analysis using Combined Deep Learning and Language Model on Real-Time Product Review

Maganti Syamala[#1], N.J.Nalini[*2]

[#1]*Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu 608002, India*
[*2]*Associate Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu 608002, India*

[#1]shyamalamaganti54@gmail.com, *[2]njncse78@gmail.com

Abstract— *Sentiment analysis is the area of study in Natural Language Processing (NLP), where it has gained its popularity in text analytics for making any kind of purchase decision. Also, there is a need for speech-based sentiment analysis in real-world applications for providing a better quality of service. But the work carried out in the speech domain has gained very less attention. So, this paper proposed a speech sentiment analysis model by considering spectrogram as an acoustic feature. The spectrogram features are trained over a deep learning model and an N-gram Language model. A combined Convolutional Neural Network (CNN) and Bi-directional-Recurrent Neural Network (Bi-RNN) architecture frameworks are implemented for acoustic modeling and a bi-gram language model to calculate the likelihood of a particular word sequence from the spoken utterance. NLP techniques like the Vader Sentiment Intensity Analyzer function is used for performing the sentiment analysis. The experimental results are analyzed in terms of Word Error Rate (WER) and Character Error Rate (CER) and proved that the proposed model holds outperforming WER and CER of 5.7% and 3 % when compared with the traditional Automatic Speech Recognition (ASR) models. The obtained sentiment analysis results are measured using correctly classified instances, precision, recall, and f1-score using various machine learning algorithms. The logistic Regression algorithm proved to achieve improved accuracy of 90% with the proposed speech sentiment analysis model.*

Keywords — *Acoustic, Character Error Rate, Convolutional Neural Network, Machine Learning, Natural language processing, Recurrent Neural Network, Speech, Spectrogram, Sentiment analysis, Word Error Rate.*

I. INTRODUCTION

The World Wide Web (WWW) evolution made an individual's life easy to make any purchase decision as WWW contains different data modalities. There is a necessity and need to process and handle this huge amount of multimedia data [26]. The evolution of machine learning, NLP, and deep learning techniques had made an individual's job easy. Sentiment analysis is a field of study in NLP that has gained its popularity in simplifying the tasks for enhancing customer service quality in recommender systems [4].

Among the vast amount of data available in WWW, the major part of sentiment analysis for opinion mining has been carried out on text. Most of the research work so far carried out had made the use of product reviews dataset for performing sentiment analysis [3]. The research work carried out on speech-based sentiment analysis is limited. Most of the research carried out on speech is ASR and Speech Emotion Recognition (SER). There is a need for speech sentiment analysis in call-centers for improving the quality of service and in YouTube reviews for analyzing the customer experience (live feedback) etc. YouTube is one of the popular social platforms to share and retrieve information. Nowadays, instead of showing interest in reading blogs and reviews, they mostly spend time on YouTube for faster information retrieval. YouTube is a place to share opinions, and one can make his/her own decision very easily and quickly. It contains a wide variety of topics related to political issues, product reviews, social issues, etc [27]. It would be a better option to perform sentiment on youtube videos as it would cover all the three variants of data like speech, the transcribed text, and image. In this paper, the work was carried on speech drawn from YouTube videos on a product review for analyzing the speech sentiment. The below Fig.1, projects the outline flow of the speech sentiment analysis model.

Even though there is a vast number of real-world applications where there is a need for speech sentiment analysis, only a few research papers have been published in this domain. As motivated by the limitations of performing sentiment analysis in the speech domain, I designed and implemented a combined deep learning and language model for speech analysis. Made the use of NLP techniques for performing sentiment analysis on the extracted text transcripts. Spectrogram speech features are trained over a combined CNN and Bi-RNN framework for deriving the acoustic features. These acoustic features are trained over a bi-gram language model for mapping character to word sequences. The proposed model achieves improved WER (%) when compared with the traditional ASR models.

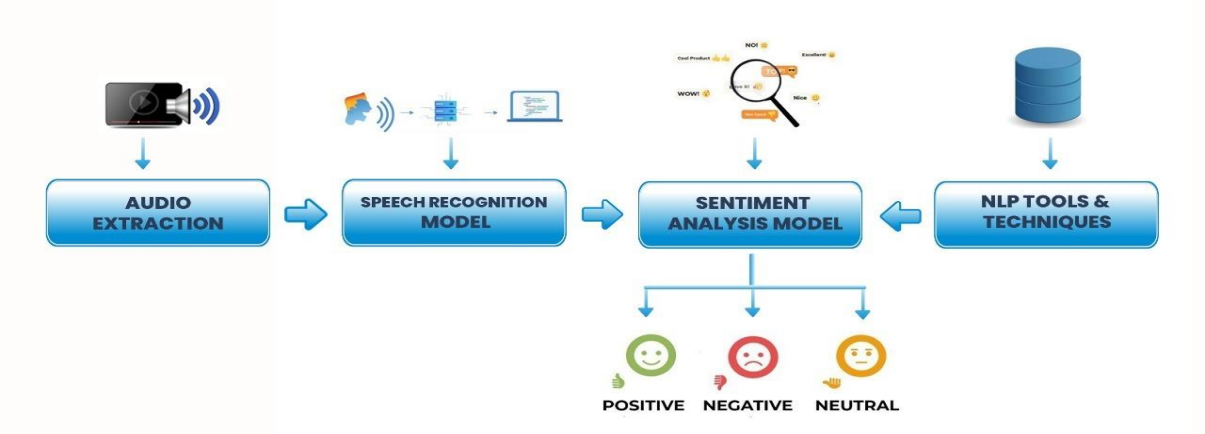

Fig. 1 Speech sentiment analysis model

II. RELATED WORK

Emotion expresses an individual feeling, which directly depends on one's vocal and oral variations. These variations may vary from person to person. So, there is a need to identify these speech features by assigning parameters with some pre-defined threshold. Many such speech features are spectral, prosodic for automatic speech recognition (ASR), and Speech emotion classification [1]. As these features directly impact the speaker, there is a need for fine-tuning of parameters for extracting the speech features in an efficient way [2]. Many studies proved that the speaker-independent models gained better accuracy and efficiency. In this paper, the literature on speech sentiment analysis is carried out to analyze how the existing studies relate acoustic features in determining the sentiment in speech.

Maghilnan S et al. proposed a speaker-specific speech sentiment analysis model. MFCC speech features are extracted and are given to a dynamic time wrapping (DTW) feature matching algorithm to obtain the text transcripts [5]. A two-phase – speaker discrimination and speech recognition architecture were implemented to map speaker I'd with the speech transcripts. The output text dialogue is given as input to the sentiment model to analyze positive, negative, and neutral sentiments in text. The accuracy of the sentiment analysis model is measured on Twitter and movie review data. The drawback with this approach is that this model can't handle a conversation and perform sentiment analysis when both the speakers speak simultaneously.

Ziquan Luo et al. proposed a sentiment analysis model by using heterogeneous speech features [6]. MFCC and other traditional acoustic features were trained over a combined deep neural architecture. A parallel combination of CNN+ LSTM is used to extract the audio sentiment vector. A Bii-LSTM network with an attention mechanism is used for feature fusion. Finally, the extracted audio sentiment vector by the trained deep learning models performs better for analyzing the sentiment.

Bryan Li et al. performed sentiment analysis on customer calls data to improve service quality [7]. Acoustic raw cepstral features at utterance level are extracted using the open Smile tool. N-gram language model is used to extract the linguistic aspects for performing sentiment analysis. Various acoustic features along with language models are compared in terms of recall to analyze the sentiment. Finally, more positivity can be analyzed from word choices, and negativity Can be analyzed from the tone of voice.

Lakshmish Kaushik et al. proposed an automatic sentiment detection system on YouTube data's natural audio streams [8]. Automatic speech recognition models (ASR) are used for decoding the text transcripts from the audio streams. Parts-of-Speech (POS) tagging, Maximum Entropy (ME) text-based models are used for sentiment detection. The experimental results are measured in terms of classification accuracy and proved that speech sentiment analysis can also achieve better accuracy, despite considering the drawbacks of poor Word Error Rate (WER).

Souraya Ezzat et al., as motivated by the use cases in call center data, proposed a sentiment analysis model on customer call recordings [9]. The proposed work mostly concentrated on using text mining techniques like text classification and text clustering to analyze the sentiment analysis. Automatic Speech Recognition Engine was used for extracting the text transcripts. WER is used as an evaluation metric to measure the performance. Bag of words, tf-idf feature extraction techniques are used, and different machine learning algorithms like SVM, KNN, decision tree, naive Bayes, and k-means are used to evaluate the classification accuracy the model.

Suresh Kumar Govindaraj et al. proposed a sentiment analysis model using acoustic and textual features on Amazon product review data [10]. The audio data has been collected from YouTube on Amazon product review. This model combined a set of acoustic features (AF) representing the audio data's emotions with the set of lexical features (LF) representing the text's sentiment by a supervised classifier to predict the customer sentiment. The experimental results were analyzed in terms of accuracy when evaluated with different LF, AF, and LF+AF combinations.

III. PROPOSED SPEECH SENTIMENT ANALYSIS MODEL

The proposed speech sentiment analysis model comprises modeling an end-to-end speech recognition

model using deep neural network architecture and a sentiment model using NLP. YouTube review dataset in its raw form is chosen for implementing the proposed model. Librispeech, an English accent speech dataset, was used for training the model. This trained dataset consists of 1000 hours of transcribed speech data. Each sample in this Librispeech dataset consists of sample rate, utterance, waveform, and necessary metadata about each speech sample. As there is a need to analyze the utterance characters in a speech signal phonetically, an acoustic spectrogram feature was extracted from the raw speech input. It is a visual representation of a signal with different spectrum frequencies that changes over time. In the proposed model, the audio wav file is transformed to Mel Spectrogram by considering a sample rate of 16000 Hz and hop size of 512. A Mel scale of size 128 is constructed to separate the frequency spectrum of equal distance. The acoustic features are trained over the deep learning model framework and an N-gram language model to obtain the text transcripts. NLP techniques are used for analyzing the sentiment on the transcribed speech.

The proposed work presents a dual approach to perform sentiment analysis on speech data. In this paper, we concentrated on how the speech can be analyzed by combining the deep learning model framework and language model for analyzing the speech to perform the sentiment analysis. Below Fig. 2, describes the design flow of the proposed speech sentiment analysis model.

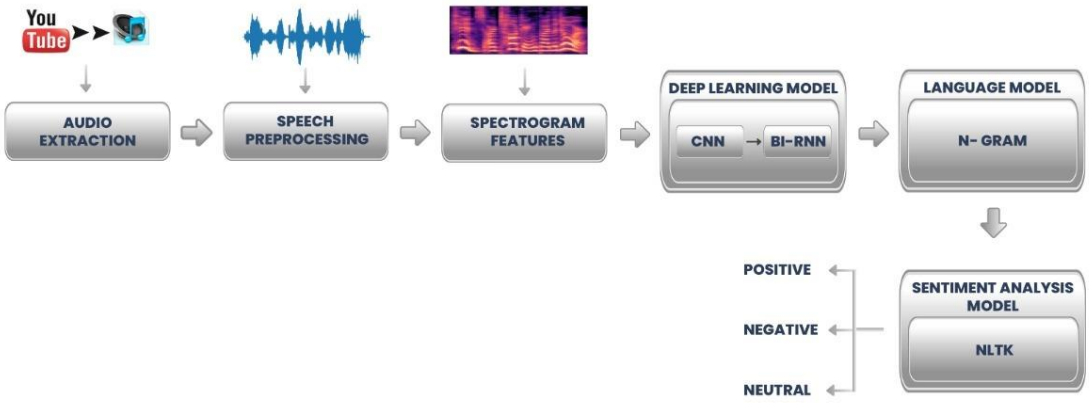

Fig. 2 Design flow of proposed speech sentiment analysis model

A. Acoustic Feature Modelling

The core of the proposed speech sentiment analysis model comprises a Convolutional Neural Network (CNN) and a Bi-directional Recurrent Neural Network (Bi-RNN). Both the networks are fully-connected to classify characters per each time step. Initially, the trained spectrograms are ingested into CNN to learn the acoustic features. The learned relevant acoustic features are fed into Bi-RNN to leverage the learned CNN acoustic features. Finally, an N-gram Language model is used to generate the English text transcriptions. WER and CER are used as model evaluation parameters, and the prediction error is handled using Connectionist Temporal Classification (CTC) loss function. Adam is used as an optimizer and scheduler during the training phase to compute gradient via backpropagation with respect to all the model parameters.

Extracting and converting the acoustic features from the given input signal to transcribed text using acoustic and language models is presented below.

Input: Spectrogram at p^{th} frequency bin at time 't' i.e.,$(f_t)^{(p)}$, Utterance 'u' and label 'l' for the given training set S_i.

$$S_i = \{(u^{(1)}, l^{(1)}), (u^{(2)}, l^2), \dots (u^{(i)}, l^{(i)})\} \quad (1)$$

Output: Probability of character set along with the transcript l_t.

$$l_t = P\left(C_t/u\right), where\ C\{a, b, c, \dots z, space, blank\} \quad (2)$$

Where $P\left(C_t/u\right)$ Denotes the probability of a character set at time 't' for the given input utterance u.

B. Pre-Emphasis

The input raw speech signal is converted into a spectrogram and the p^{th} frequency bin in each frame of the audio segment $(f_t)^{(p)}$ Is given as feature input to the CNN model.

for each utterance $u^{(i)}$

$$u^{(i)} = len\left(T^{(i)}\right) \quad (3)$$

$$T(i) = u_t{}^{(i)} = \{u_t{}^{(1)}, u_t{}^{(2)}, \dots \dots, u_t{}^{(i)}\} \quad (4)$$

$$u_{t,p}{}^{(i)} = (f_t)^{(p)} \quad (5)$$

C. Deep Learning Model

The deep learning model is designed using CNN and RNN. The total number of hidden layers considered for designing the model architecture is 5. The CNN model is a non-recurrent neural network with the first three hidden layers, and the next 2 hidden layers are RNN.

Input: Utterance u, hidden units (h) at layer (l) h_l=h_0, h_1, h_2, h_3, h_4

CNN = h_0, h_1,h_2
RNN = h_3, h_4
Output: SoftMax layer h_5

1) CNN Model

The CNN architecture designed is a fully connected layer architecture with 3 hidden layers h_0, h_1, h_2. The feature maps from the input spectrogram are feed-in sequence to CNN. The convolution and max-pooling layers of CNN perform necessary operations to generate a sequence of weights (activation functions). The output of each layer at time 't' depends on the input spectrogram feature map u_t and context frames C_t on each side at the time 't.' The output from the last hidden layer is feed as input into Bi-RNN.

for each hidden layer in CNN, h_0, h_1,h_2

$$\left(h_{0,1,2}\right)_t = u^{(i)}{}_{t,p} + C_t \quad (6)$$

$$\left(h_{0,1,2}\right)_t = g\left(W_l\left(h_{(l-1)}\right)_t + b_l\right) \quad (7)$$

2) Bi-RNN Model

The Bi-RNN model designed is a bi-directional recurrent neural network with 2 hidden layers h_3, h_4. It uses a forward recurrence $h^{(f)}$ and backward recurrence $h^{(b)}$ to map each utterance u_i in a forward and reverse direction. W_l and $b_l$are the weight matrix and bias parameters at each layerh_l.

for each hidden layer in Bi-RNN, h_3, h_4

$$\left(h_{3,4}\right)_t = b_f + b_b \quad (8)$$

$$\left(h_{3,4}\right)^f{}_t = g\left(W^{(4)} h_t{}^{(3)} + W_r{}^{(f)} h_{l-1}{}^{(f)} + b^{(4)}\right) \quad (9)$$

$$\left(h_{3,4}\right)^b{}_t = g\left(W^{(4)} h_t{}^{(3)} + W_r{}^{(b)} h_{l+1}{}^{(b)} + b^{(4)}\right) \quad (10)$$

for each utterance u_i

$$h^{(f)} = t\left(1\ to\ T^{(i)}\right) \quad (11)$$

$$h^{(b)} = t\left(T^{(i)\ to\ 1}\right) \quad (12)$$

3) SoftMax Layer

The final output layer uses a SoftMax function to display the predicted output $\mathrm{P}\left(\mathrm{C_p}/\mathrm{u_i}\right)_\mathrm{t}$ The probabilities of each character' n' in the alphabet at every time slice 't.'

$$(h_5)_t = g\left(W^{(5)} h_t{}^{(4)} + b^{(5)}\right) \quad (13)$$

$$h_t{}^{(4)} = \left(h_t{}^{(f)} + h_t{}^{(b)}\right) \quad (14)$$

$$P\left(C_p/u_i\right)_t = exp\left(W_n{}^{(6)} h_t{}^{(5)} + b_n{}^{(6)}\right) \Big/ \sum_j exp\left(W_j{}^{(6)} h_t{}^{(5)} + b_j{}^{(6)}\right) \quad (15)$$

D. Language Model (LM)

The proposed model uses an N-gram language model to compute and generate the text transcripts. Based on the chain rule $P(M_s/u)$, used a bi-gram language model to compute the joint probability to retrieve the occurrence of a maximum character sequence in the utterance.

Input: Output from RNN model, character level transcripts.

Output: Maximum Sequence of characters that achieves the combined objective.

for each $P\left(C_p/u_i\right)_t$ from Bi-RNN

$$Max\_Seq(M_s) = \log(P(M_s/u) + \alpha \log\left(P_{lm}(M_s)\right) + \beta\ word\_count(C) \quad (16)$$

where α denotes the language model constraint, and β denotes the length of the sentence.

$$P(M_s/u) = \prod_i P(u_1 | u_1 u_2, \ldots . u_{i-1}) \quad (17)$$

E. Sentiment analysis Model

The text transcripts are trained over a pre-trained 'Vader' sentiment model imported from the NLTK python package to perform sentiment analysis. SentimentIntensityAnalyzer() function imported from nltk.sentiment.Vader package was used to analyze the sentiment in the transcribed text in terms of positive, negative, and neutral sentiment.

IV. RESULTS AND DISCUSSIONS

Experimental results are implemented using python programming language on real-time product review data. As the proposed work focuses on speech sentiment analysis, I collected the Samsung mobile reviews data as a dataset from YouTube. Product reviews on Samsung M31 have been downloaded from YouTube. YouTube, a social platform where people share their life experiences in reviews, has a natural, spontaneous speaking style. Like the way the speaker speaks directly impacts describing the accuracy of the model, it made me motivate and download the dataset from YouTube. In total, 40 YouTube reviews of about size 90 KB and duration of about 12 minutes having the strong presence of subjectivity, positivity, and negativity are randomly collected and converted to .wav files for performing speech sentiment analysis.

Word Error Rate (WER) and Character Error Rate (CER) are the two-evaluation metrics used for evaluating the performance of the proposed speech recognition model. Levenshtein distance, a string similarity metric, can measure the differences between two string sequences in insertions, deletions, and substitutions. WER and CER are calculated using this Levenshtein distance, and the way these two metrics are calculated are detailed below.

- Word Error Rate (WER): WER evaluation metric compares the reference text of our input audio with the text transcribed from our model at the word level.

$$WER = (WS + WD + WI)/WR \quad (18)$$

Where WS denotes the number of words substituted in the hypothesis test.

WD denotes the number of words deleted in the hypothesis test.

WI denotes the number of words inserted in the hypothesis test.

WR denotes the number of words in the reference text.

- Character Error Rate (CER): CER evaluation metric compares the reference text of our input audio with the text transcribed from our model at the character level.

$$CER = (CS + CD + CI)/CR \quad (19)$$

Where CS denotes the number of characters substituted in the hypothesis test.

CD denotes the number of characters deleted in the hypothesis test.

CI denotes the number of characters inserted in the hypothesis test.

CR denotes the number of characters in the reference text.

The following analysis has been carried out to compare our proposed model with traditional speech and sentiment analysis approaches.

- The proposed speech sentiment analysis model was made a comparison with the traditional ASR models for speech analysis.
- The input spectrogram acoustic features considered in the proposed model, along with the deep learning model for acoustic feature extraction, compare with the traditional acoustic feature models.
- The importance of the proposed hybrid CNN+Bi-RNN framework is analyzed.
- The impact and the use of the N-gram language model were compared with the traditional Hidden Markov model (HMM) and Listen, Attend, Spell (LAS)language models.
- Model performance was evaluated using WER and CER as evaluation metrics [25] between the traditional ASR models and the proposed model.
- Various performance measurement metrics like accuracy, precision, recall and f1-score are computed to analyze the impact of the proposed speech sentiment analysis model using different machine learning algorithms.

Below Table 1 compares different speech recognition models implemented on the trained Librispeech dataset. The combination of acoustic features, deep learning models, and language models is compared to evaluate the model performance in terms of Word Error Rate (WER).

Table I: Model Performance Comparison in terms of WER (%)

Reference	Acoustic Feature (AF)	Deep Learning Model (DM)	Language Model (LM)	WER (%)
Tom ko et al., [11]	MFCC	DNN	4-gram	12.51
VitaliyLiptchinsky et.al, [12]	Log-Mel Filter Banks	Gated ConvNet	4-gram	14.5
Vassil Panayotovet.al, [13]	fMLLR	HMM-DNN	4-gram	13.97
Dario Amodei et.al, [14]	Log Spectrograms	2D-CNN	N-gram	12.73
Albert Zeyer et al., [15]	Audio Features	LSTM	LSTM	12.76
Neil Zeghidour et al., [16]	Mel-Filter Banks	Conv AM	Conv LM	10.47
Christoph Luscher et.al, [17]	MFCC	DNN / HMM	LSTM	9.3
Kyu J. Han et al., [18]	MFCC	TDNN – LSTM + CNN-BLSTM + Dense TDNN-LSTM	4-gram	7.64
Duc Le et al., [19]	Audio Features	LC – BLSTM AM	4-gram	7.6
Proposed Model	**Spectrogram**	**CNN + Bi-RNN**	**2-gram**	**5.7**

The proposed model by combining spectrogram acoustic features with a CNN + Bi-RNN framework and bi-gram language models yields outperforming WER compared with other combinations in the state-of-art methods. The results are plotted in the below Fig. 3.

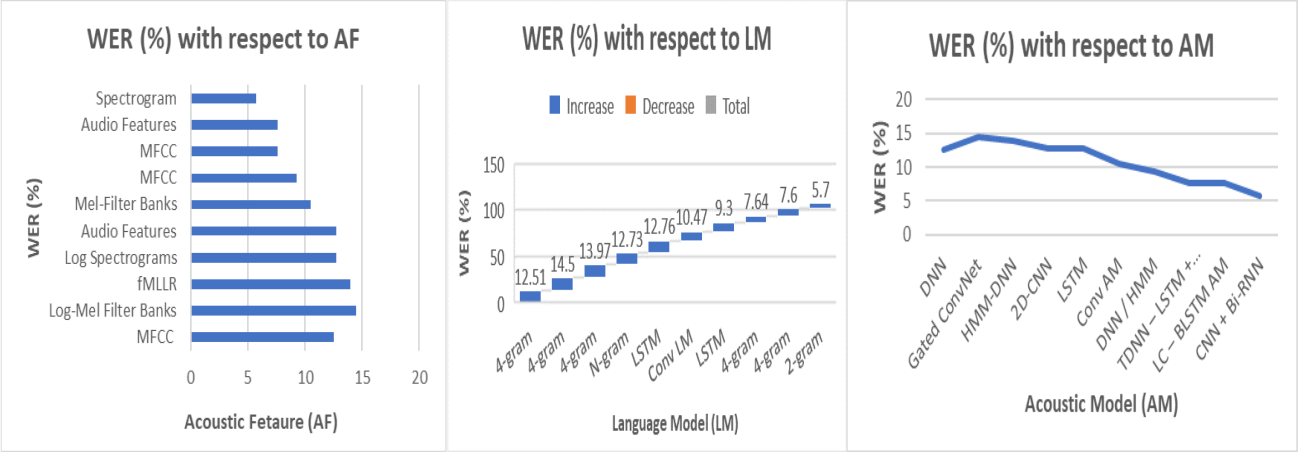

Fig. 3 Model Performance Comparison in terms of WER (%)

The percentage values in Table 2 depicts the WER of speech scripts when translated using online speech API's.

Table II: Comparison of Online Speech-text API's in terms of WER (%)

Online Speech-to-Text API	WER (%)		
	Pre-processed (Clean)	Un-Filtered (Noisy)	Combination
Apple Dictation API	14	44	27
Bing-Speech API	12	36	22
Google API	7	30	17
Wit.ai API	8	35	19

A comparison was made between different online speech-text API's like Apple Dictation, Bing-Speech, Google, Wit.ai. The drawbacks with these online API's are that they can process audio for less than 1 minute conversion and are payable after limited use. Among the compared online speech-text APIs, Google API yields good results with less WER, and the results are shown in the below Fig. 4.

The model accuracy gets improved from the literature by the use of recurrent neural networks as an acoustic model. So, made a comparison in terms of Character Error Rate (CER) on the train and tested Librispeech data with Deep Neural Network (DNN), Forward Recurrent Neural Network (RNN), and Bi-directional Recurrent Neural Network (Bi-RNN) in Table 3.

The values are measured among three different types of speech scripts: clean, noisy, and both.

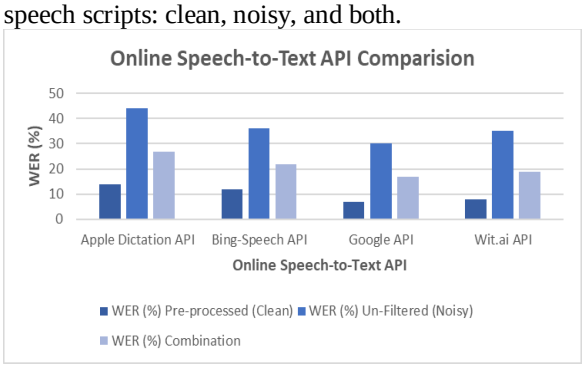

Fig. 4 Comparison of Online Speech-text API's in terms of WER (%)

Table III: Comparison with different variants of Recurrent Neural Network in terms of CER (%)

Deep Learning Model	CER (%) on Trained model	CER (%) on the Test model
Deep Neural Network	4	22
Forward Recurrent Neural Network	4	14
Bi-directional Recurrent Neural Network	3	11

In Fig. 5, different variants of RNN are compared, and among all the three compared variants of the recurrent neural network, Bi-directional Recurrent Neural Network proved to achieve less CER (%).

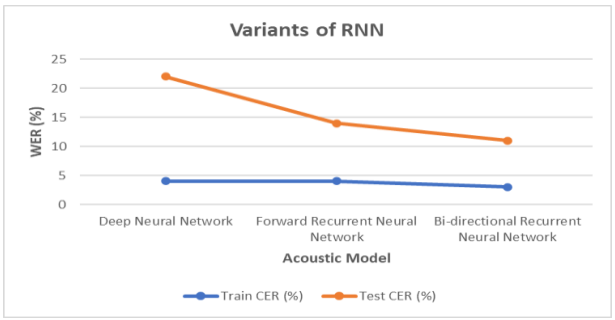

Fig. 5 Comparison with different variants of Recurrent Neural Network in terms of CER (%)

A comparison was made to analyze the speech recognition model's performance with respect to WER between the models with language model (LM) and

without LM. The results are depicted in Table 4 and the LM on clean and noisy trained Librispeech data.

Table IV: Comparison with LM and without LM in terms of WER (%)

Language Model	Reference	WER (%)			
		No Language Model		With Language Model	
		Clean	Noisy	Clean	Noisy
HMM	Han et.al, [20]	-		4	9
	Yang et.al, [21]	-		3	8
CTC	Liptchinsky et.al, [12]	7	21	5	15
	Li et.al, [22]	4	12	3	9
LAS	Zeyer et.al, [23]	5	15	4	13
	Irie et.al, [24]	5	13	4	10

Sentiment analysis was carried out using the Vader Sentiment NLP technique on the YouTube text transcript obtained from the Speech Recognition model. Machine learning algorithms like Support vector machine, Naïve Bayes, Logistic regression, Decision tree, K-Nearest Neighbor, and Random Forest are used to analyze the performance of the proposed speech sentiment analysis model. The performance was analyzed in terms of correctly classified instances, accuracy, precision, recall, and f1-score measures. The logistic Regression algorithm proved to achieve better accuracy for our proposed model compared to the other algorithms shown in Table 5 and Fig. 6.

Table V: Performance measures comparison with different machine learning algorithms

Machine Learning Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Support Vector Machine	87	89	93	90
Naïve Bayes	84	85	90	89
Logistic Regression	90	92	95	94
Decision Tree	78	86	86	86
K-Nearest Neighbor	65	86	70	74
Random Forest	80	88	87	87

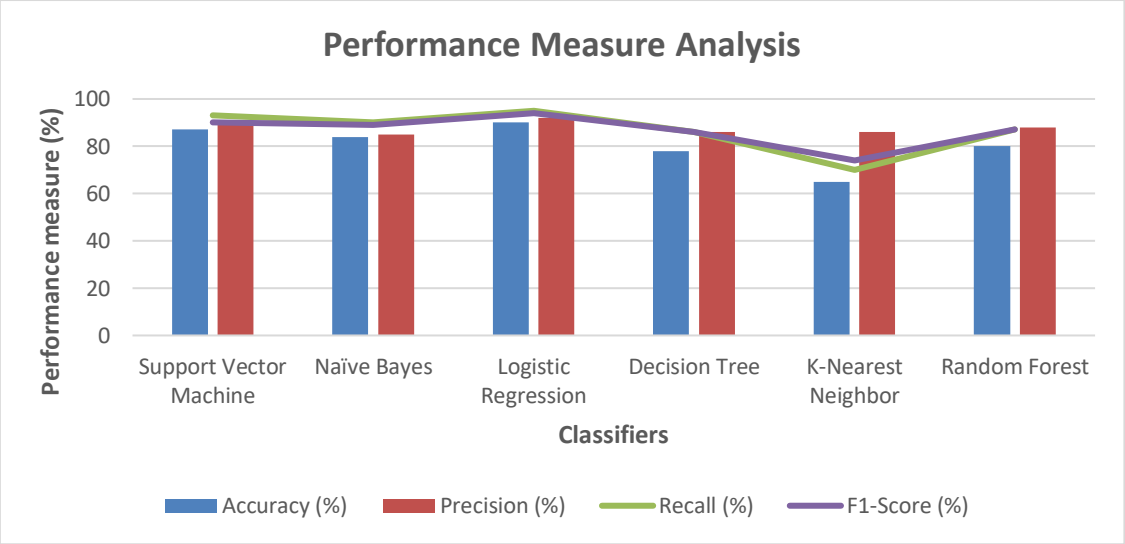

Fig. 6 Performance measures comparison with different machine learning algorithms

V. CONCLUSION

In this paper, an improved speech sentiment analysis model was implemented on real-time product review speech data drawn from the YouTube platform. As most of the research carried on sentiment analysis in the speech domain uses online speech recognition API's, it has several constraints and drawbacks. So, proposed an improved end-to-end speech recognition model by implementing a combined CNN and Bi-RNN deep learning model by training speech spectrogram features. A bi-gram language model maps the obtained character features from the deep

learning model into respective word sequences. The proposed model is independent of the input speech file's duration and can predict the correct word sequences with respect to the spoken utterances. WER and CER are used as the evaluation metrics to measure the error rate of our proposed speech recognition model. The experimental results obtained proved to achieve improved WER and CER when compared with the traditional ASR models. Sentiment analysis was performed using NLP techniques on the transcribed text scripts and achieved 90% accuracy when validated using the Logistic Regression machine learning algorithm.

VI. FUTURE WORK

This paper mostly proves to achieve improved speech to text classification accuracy for sentiment analysis. In the future, we would like to enhance the proposed speech analysis model by combining with different text analytics and feature extraction techniques for implementing the improved aspect-based speech sentiment analysis model.

REFERENCES

[1] M.S. Hossain, G. Muhammad.,Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data, Information Fusion, (2018) 1-24.

[2] Y. Zeng, H. Mao, D. Peng, Z. Yi.,Spectrogram based multi-task audio classification. Multimed. Tools Appl, 78(2019) 3705–3722, 2019.

[3] M. Syamala, and N.J. Nalini.,A filter-based improved decision tree sentiment classification model for real-time amazon product review data., International Journal of Intelligent Engineering and Systems, 13(1)(2020) 191-202.

[4] M. Syamala, and N.J. Nalini.,A deep analysis on aspect-based sentiment text classification approaches, International Journal of Advanced Trends in Computer Science and Engineering, 8(5) (2019) 1795-1801.

[5] S. Maghilnan and M. R. Kumar.,Sentiment analysis on speaker-specific speech data, in Proceedings of International Conference on Intelligent Computing and Control (I2C2), (2017) 1-5.

[6] Z. Luo, H. Xu, and F. Chen.,Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network, AffCon@AAAI, (2019).

[7] B. Li, D. Dimitriadis and A. Stolcke.,Acoustic and Lexical Sentiment Analysis for Customer Service Calls,in Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (2019) 5876-5880.

[8] L. Kaushik, A. Sangwan, and J. H. L. Hansen.,Sentiment extraction from natural audio streams, in Proceedings of International Conference on Acoustics, Speech and Signal Processing, 8485-8489, (2013).

[9] S. Ezzat, N. E. Gayar, and. M. Ghanem.,Sentiment Analysis of Call Centre Audio Conversations using Text Classification, International Journal of Computer Information Systems and Industrial Management Applications, 4(2012) 619 -627.

[10] S. Govindaraj, and K. Gopalakrishnan, "Intensified Sentiment Analysis of Customer Product Reviews Using Acoustic and Textual Features, ETRI Journal, 38 (3)(2016) 494-501.

[11] T. Ko, V. Peddinti, D. Povey, S. Khudanpur, "Audio Augmentation for Speech Recognition. in Proceedings of INTERSPEECH, (2015).

[12] V. Liptchinsky, G. Synnaeve, R. Collobert.,Letter-Based Speech Recognition with Gated ConvNets., ArXiv abs/1712.09444, 2017.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur., Librispeech: An ASR corpus based on public domain audiobooks, Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2015) 5206-5210.

[14] Dario Amodei et al.,Deep speech 2: end-to-end speech recognition in English and mandarin, in Proceedings of the 33rd International Conference on International Conference on Machine Learning, 48(2016) 173–182.

[15] A. Zeyer, K. Irie, R. Schluter, and H. Ney.,Improved training of end-to-end attention models for speech recognition., in Proceedings of INTERSPEECH, (2018).

[16] N. Zeghidour, Q. Xu, V. Liptchinsky, N.Usunier, G. Synnaeve, and R. Collobert.,Fully Convolutional Speech Recognition., (2018).

[17] Ch. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney.,RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation,(2019).

[18] K. Han, A. Chandrashekaran, J. Kim, and I. Lane.,The CAPIO 2017 Conversational Speech Recognition System, (2017).

[19] D. Le, Duc, X. Zhang, W. Zheng, Ch. Fugen, G. Zweig, and M. Seltzer.,From Senones to Chenones: Tied Context-Dependent Graphemes for Hybrid Speech Recognition, in Proceedings of International Conference of Automatic Speech Recognition and Understanding Workshop (ASRU)IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 457-464, (2019).

[20] K. J. Han, A. Chandrashekaran, J. Kim, and I. Lane.,The CAPIO 2017 Conversational Speech Recognition System, in arXiv, (2017).

[21] X. Yang, J. Li, and X. Zhou.,A novel pyramidal-FSMN architecture with lattice-free MMI for speech recognition, in arXiv, (2018).

[22] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, Jasper: An End-to-End Convolutional Neural Acoustic Model, in arXiv, (2019).

[23] A. Zeyer, A. Merboldt, R. Schluter, and H. Ney.,A comprehend-¨sive analysis on attention models., in NIPS: Workshop IRASL,(2018).

[24] K. Irie, R. Prabhavalkar, A. Kannan, A. Bruguier, D. Rybach, and P. Nguyen.,Model Unit Exploration for Sequence-to-Sequence Speech Recognition.,in arXiv, (2019).

[25] S. Benkerzaz, Y. Elmir and A. Dennai.,A Study on Automatic Speech Recognition, Journal of Information Technology Review, 10(2019) 77-85.

[26] L. P. Maguluri, and R. Ragupathy.,Comparative Analysis of Companies Stock Price Prediction Using Time Series Algorithm, International Journal of Engineering Trends and Technology 68.11(2020) 9-15.

[27] Kanusu Srinivasa Rao,Mandapati Sridhar., Sustainable Development of Green Communication through Threshold Visual Cryptography Schemes Using a Population Based Incremental Learning Algorithm, Journal of Green Engineering,11(1)(2020) 608-624.