International Advance Journal of Engineering Research (IAJER)
Volume 3, Issue 3 (March- 2020), PP 40-47
ISSN: 2360-819X
www.iajer.com

Research Paper **Open Access**

Novel Approach on Audio to text Sentiment Analysis on Product Reviews

[1]Jayashri Patil, [2]Priti Subramanium
[1] PG Student, [2] Assistant Professor
[1,2]Computer Science and Engineering,
[1,2]Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, India.

ABSTRACT:- One Now a days to human-machine interaction is estimating the speaker's emotion is a challenge. The need is more accurate information about consumer choices increasing interest in high-level analysis of online media perspective. Usually in emotion classification, researchers consider the acoustic features alone. For strong emotions like anger and surprise, the acoustic features pitch and energy are both high. In such cases, it is very difficult to predict the emotions correctly using acoustic features alone. But, if we classify speech solely on its textual component, The uniqueness in this approach is the generation of text sentiments, audio sentiments and blend them to obtain better accuracy**.** In this paper, I have proposed an approach for emotion recognition based on both speech and media content. Most of the existing approaches to sentiment analysis focus on audio and text sentiment. This novel approach is the generation of text sentiments, audio sentiments and blend them to obtain better accuracy**.**

Keywords:- Sentiments, Features, Natural Language Programming, Hybrid, Accuracy.

I. INTRODUCTION

Emotion recognition plays a major role in making the human-machine interactions more natural. In spite of the different techniques to boost machine intelligence, machines are still not able to make out human emotions and expressions correctly. Emotion recognition automatically identifies the emotional state of the human from his or her speech. One of the greatest challenges in speech technology is evaluating the speaker's emotion. Most of the existing approaches focus either on audio or text features. In this work, I have proposed a novel approach for emotion classification of audio conversation based on both speech and text. The novelty in this approach is in the selection of features and the generation of a single feature vector for classification. My main intention is to increase the accuracy of emotion classification of speech by considering both audio and text features.

An everyday enormous amount of data is created from social networks, blogs, and other media and reused into the World Wide Web. This huge data contains very crucial opinion related information that can be used to benefit businesses and other aspects of the commercial and marketing industries. Manual tracking and extraction of this useful information are impossible; thus, sentiment analysis is required. Sentiment Analysis is extracting sentiments or opinions from reviews expressed by users over a particular subject, area or product using online data. It is an application of natural language processing, computational linguistics, and text analytics to identify subjective information from source data. It divides the sentiment into categories like Positive, Neutral and Negative sentiments. Thus, it determines the general attitude of the speaker or a writer with respect to the context of the topic.

Natural language processing (NLP) is the field which comes under artificial intelligence dealing with our most ubiquitous product: the context in emails, web pages, tweets, product descriptions, newspaper stories, social media, and scientific articles in thousands of languages and varieties. Successful natural language processing applications are an important of our everyday experience, spelling and grammar correction in word processors, machine translation on the web, email spam detection, automatic question answering, detecting people's opinions about products or services, extracting appointments to your email.

The sentiment analysis result is based on the accuracy of the machine learning model. To increase the accuracy of the model I will be considering both audios of the reviewer and its textual context. I am going to increase the accuracy of the model as both speech and text, sentiments are considered rather than going for only text or speech.

II. LITERATURE SURVEY

Jasmine Bhaskar et. al. were used two datasets for their experiments. In the first stage the dataset consisted of news headlines which were taken from SemEval-2007 and Google. As standard data was unavailable, the waveform of audio of the corresponding dataset was developed. For these 360 vectors, each with 11 features was used to train the SVM classifier [1]. In the second experiment, the accuracy of text emotional Classifications was tested. Again, 360 vectors each with 85 features representing the emotion words in the dataset were generated and these vectors were used to train the SVM classifier. Common features such as pitch, energy, formants, intensity, and ZCR (Zero Crossing Rate). In this work, they have extracted the formants, zero crossing rate, and sound intensity by using MATLAB and fundamental frequency (F0), energy using praat.

In their experiment, the (NLTK) in python is used for pre- processing. English lexical database of WordNet is used for text classification. It is a well-known and popular lexical resource, it identifies as the emotions that the words convey. The emotions classification based on Vector representation of the document and emotion classification using SVM shown in Table 2.1. In their hybrid approach, they used multi-class SVM for emotion Classifications.

Table. 2.1 Accuracy of Classification Approach [1]

Classification Approach	Accuracy
Speech emotion Classification	**57.1 %**
Text emotion classification	**76 %**
Hybrid approach	**90**

The accuracy of the above differential classification approach are given in Figure. 1. The hybrid approach achieved the highest accuracy and shows better improvement compared to others. The text mining did well compare to speech as most of the emotion words in the dataset are considered in generating the feature vector.

Table 2.2 Shows the confusion matrix for the hybrid approach. The confusion matrix shows the accuracy for classified emotion count and misclassified emotion count. Each emotion has less misclassification in the hybrid approach. Anger, fear, disgust, surprise, happy and sad emotion have seven, five, three, three, two, six misclassifications respectively.

Table 2.2 Confusion matrix for the hybrid approach

Class	Happy	Sad	Fear	Disgust	Surprise	Anger
Happy	48	0	0	0	2	**0**
Sad	1	44	0	3	0	**2**
Fear	1	3	45	1	0	**0**
Disgust	1	0	2	46	0	**0**
Surprise	3	0	0	0	47	**0**
Anger	**1**	**1**	**2**	**0**	**3**	**43**

Layla Hamieh et al, have collect data set contains 350 different movie quotes along with their corresponding texts extracted online. The movies from which the quotes were extracted were chosen on the basis of making their dataset diverse enough to have a very different emotional tag equally expressed.

In first step , they have processed the text and extracted the features for classification. After that, they have processed the data using the Stanford part of speech tagger. The audio features were extracted using an English lexical database from WordNet as an act to obtain emotional tags. The audio is pre-processed and correspondingly features were extracted. Data were pre-processed based on the frequency of human voice which ranges between 300 and 3400 Hz. Consequently, to only keep data corresponding to the human voice and remove all irrelevant noise data, a bandpass filter with cut-off frequencies [300 Hz, 3400 Hz] was applied to each audio segment. Five different types of speech-related features were considered in order to diversify the features spectral:

1. Audio Spectrum roll-off is the frequency below which 85
2. Audio Spectrum centroid is the center of mass of the spectrum. It is the mean of the frequencies of the signal weighted by their magnitudes, determined using a Fourier transform.
3. Mel-Frequency Cepstral Coefficients (MFCCs) which express the audio signal on a Mel-Frequency Scale (linear below 1000Hz and logarithmic above 1000Hz). These coefficients help in identifying

phonetic characteristics of speech.

4. Zero Crossing: In a given period, the number of times the time domain signal crosses zero is the zero crossings.
5. Log-attack time is the time taken by a signal to reach its maximum amplitude from a minimum threshold time.

The data vectors collected from speech and text separately inputted to the SVM classifier. 10-fold cross-validation was used to test the accuracy of each alone. These tests were used as a base to compare with the proposed fusion method. Fusion Technique: In their method, audio and textual features were consolidated into a single feature vector before being input to the classifier (SVM).

In order to analysis of the algorithm accuracy, three experiments are conducted for detecting which emotional classification algorithm provided the highest accuracy. In the rest experiment, they tested for speech Classification. First, they have converted all media files to .wav (Waveform Audio File Format) since they were in better quality and it was easier to interact with MATLAB. The SVM classifier was trained with the collected 262 vectors each with 183 attributes. The 183 attributes were the speech extracted features described in section IV. In the second experiment, they have calculated the accuracy of a text emotional Classification algorithm. Again, the SVM classifier was trained with 262 vectors each with 5 attributes representing the score of each emotion. In the third experiment, they have calculated the accuracy of the fusion algorithm. This time, the SVM classifier was trained with 262 vector searches with 188 attributes. Both speech and text extracted features were concatenated and used as an input to the classifier. In the three experiments, the Classification was done with 10-fold cross-validation. The results of the experiments are shown in Table2.3

Table.2. 3 Simulation results

Algorithm	Accuracy
Speech emotional classification (180 features)	**45.42 %**
Text emotional classification (5 features)	**26.36 %**
Fusion approach (180+5 features)	**46.57 %**
Speech emotional classification (50 features)	**35.88 %**
Fusion approach (50+5 features)	**31.30 %**

III. SCOPE AND OBJECTIVES FOR PROPOSED SYSTEM:

Most of the existing approaches focus either on audio or text features. As per the analysis, I have found out that accuracy is less than the model having a hybrid approach. Various online text reviews can be used to determine the output sentiment. Also, various social media audio input is also significantly been used to determine output sentiment. Some of the existing star-based rating systems are using text context as input to determine the rating. The existing approach present consist only consider either text features or audio features to determine the sentiment. I have used both features to be considered as single feature vector to increase the accuracy of the system.

The existing system have considered numerous features to classify the emotion in which many emotion are not that significant, therefore I have considered only few audio and text feature in increasing the performance of the system.

To develop an application which works on both text and speech sentiment analysis for more accurate product reviews of customers. Comparison of accuracy of individual sentiments and hybrid sentiments can be achieved. To help the business organization to improve the overall quality of their service by providing them details of the analysis. Forecasting can be achieved for businesses by knowing their customer's mindset.

In this paper, I have proposed a novel approach for emotion classification of audio conversation based on both speech and text. The novelty in this approach is in the choice of features and the generation of a text and audio sentiment for classification. My main intention is to increase the accuracy of emotion classification of speech by considering both audio and text features.

IV. SYSTEM REVIEW

These are untruthful reviews that are not based on the Reviewer's genuine experiences of using the products or services, but with hidden motives. They often contain undeserving positive opinions about some target entities (products or services) in order to promote the entities and/or unjust or false negative opinions about some other entities in order to damage their reputations. First, fake reviewers actually like to use I, myself, mine, etc., to give readers the impression that their reviews express their true experiences. Second, fake reviews are not necessarily traditional lies. For example, one launches a product and pretended to be a user of that product and gives a review to promote the product. The review might be the true feeling of the author. Furthermore, many fake reviewers might have never used the reviewed products/services, but simply tried to give positive or negative reviews about something that they do not know. They are not lying about any facts they know or their true feelings. Fake reviews may be stated by many types of people, e.g., friends and family, company employees, competitors, businesses that provide fake review services, and even genuine customers.

Audio and text file will be static

In this system, the audio file and text file should be available if you want to correct answer. The audio-related text file should be available in the system. This project is not working if we have a new audio file and we don’t have text file related to it.

Sarcasm

Sarcasm is generally characterized as satirical with that is intended to insult, mock, or amuse. Sarcasm can be manifested in many different ways, but recognizing sarcasm is important for natural language processing to avoid misinterpreting sarcastic statements as literal. For example, sentiment analysis can be easily misled by the presence of words that have a strong polarity but are used as the opposite polarity was intended. The distinctive quality of sarcasm is present in the spoken word and manifested chief by vocal inflections. The sarcastic content of a statement will be dependent upon the context in which it appears. First, situations may be ironical, but only people can be sarcastic. Second, people may be ironical unintentionally, but sarcasm requires intention.

V. METHODOLOGY

Proposed System architecture

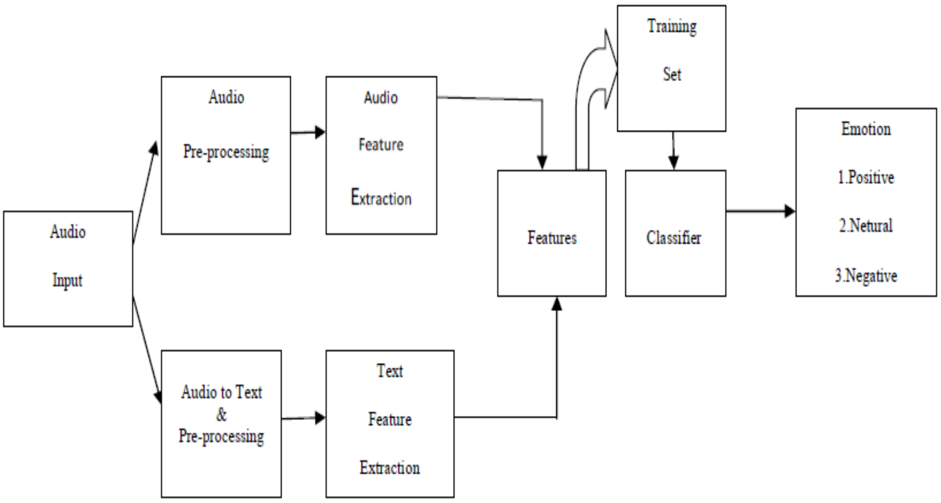


Figure. 5.1 Novel Approach on Audio to text Sentiment Analysis on Product Reviews

In this paper, I have proposed a novel approach for emotion classification of audio conversation based on both speech and text. The novelty in this approach is in the choice of features and the generation of a text and audio sentiment for classification. My main intention is to increase the accuracy of emotion classification of speech by considering both audio and text features.

Reviews:

Input given to sentiment analysis is an audio file. This audio file is of product reviews. These files are collected from review sites as well as it can be given by users as input.

Processing:

In processing, shown in Figure 5.1 the audio file is given to two different systems. One is given to text to speech converter and other to the audio algorithm. The text to speech converter converts the audio file into a text file. This converted text file is then given to text algorithms. First, it is given to emotion Classification algorithm and next to text feature extraction algorithm. While the output of the audio algorithm is given to audio feature extraction. The features extracted from both text and audio feature extractor gets combine into one in a single feature vector. The output of a single feature vector is given to model which gives the result.

Classification:

Review of audio file is classified as a positive, neutral and negative using model. The model is based on machine learning classifier. Algorithms are trained on sample labeled review text to build a model. A trained model of the classifier is then used for categorization of new test reviews.

Summarization:

In this method, the sentiment scores for every aspect is aggregated in order to represent a summary. This data is represented to the user as a class. The Summarized review information will assist users in decision making about Product.

VI. EXPERIMENTAL RESULT

I have considered one database set for the experiment. It consist of audio files of .wav format converted from using youtube audio reviews.I have conducted various experiment and develop an application in .Net Windows forms. The initial application is looking like below figure 6.1 to Figure 6.4.

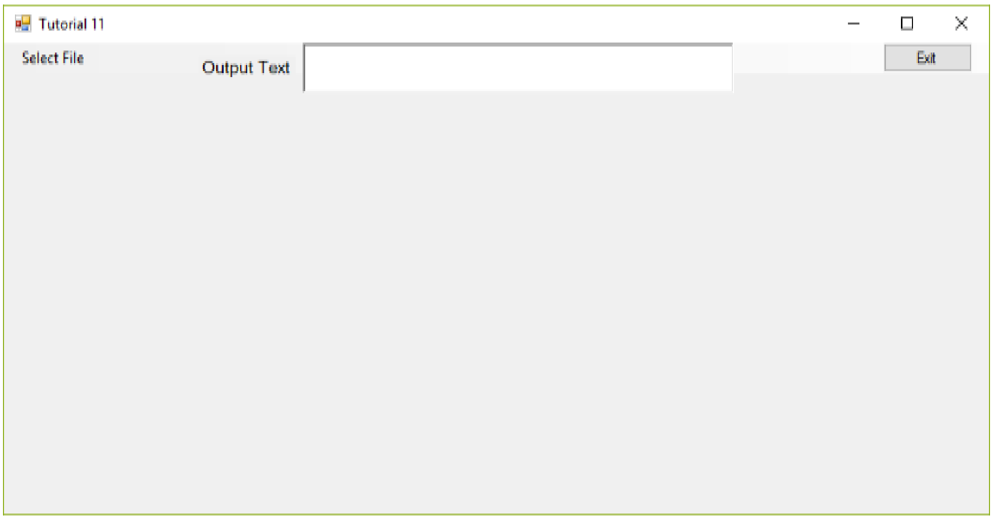
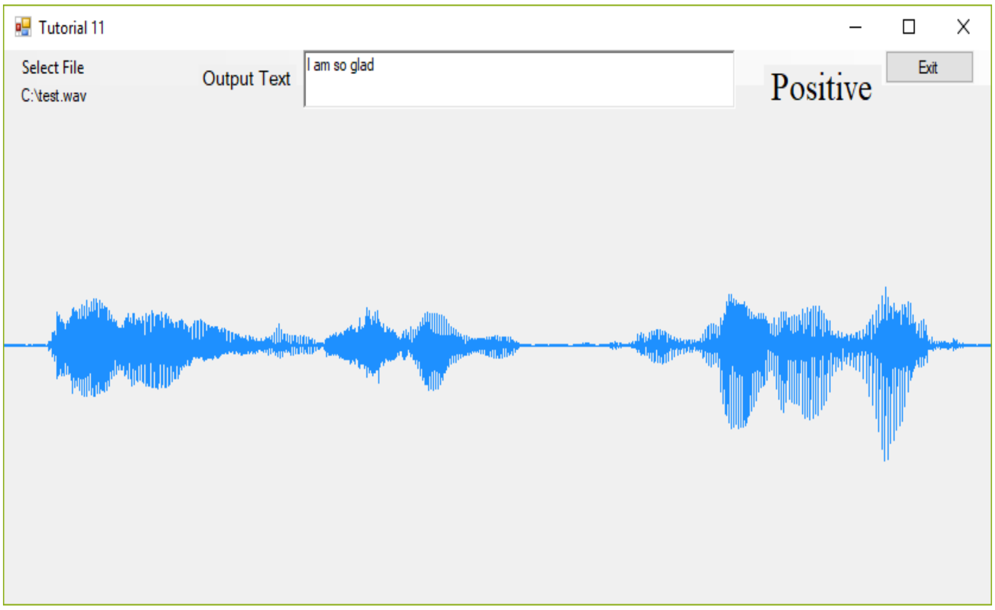


Figure 6.1 Input

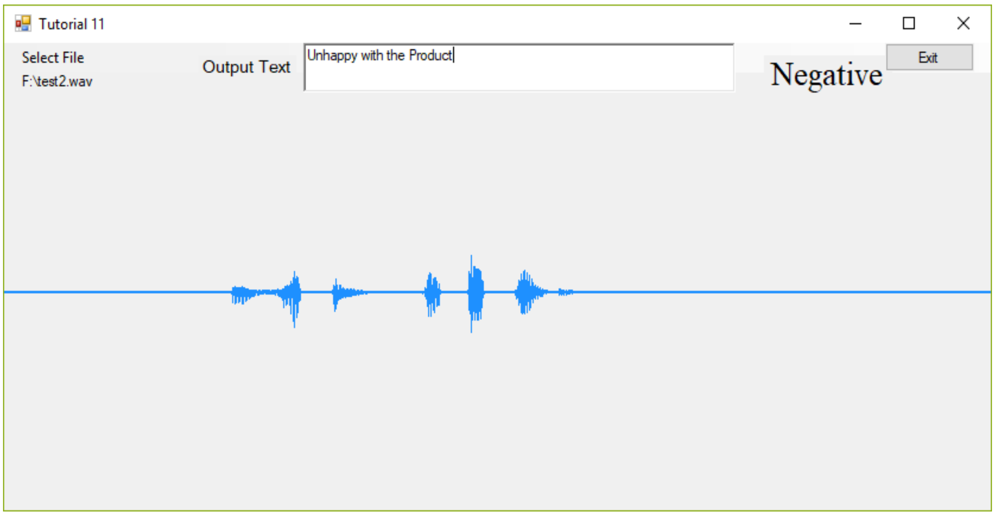


C:\test.wav

I am so glad

Positive

Figure. 6.2 Output for Positive review



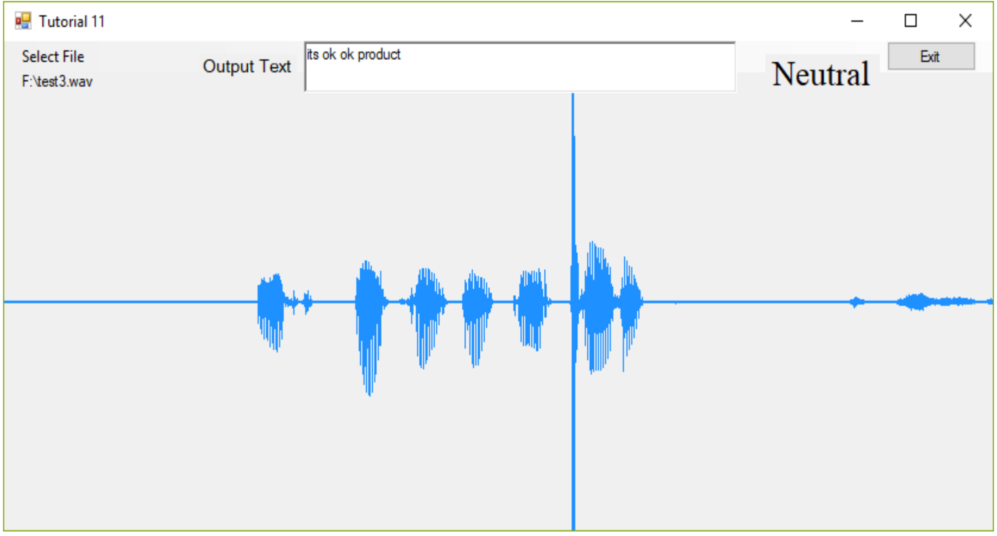
F:\test2.wav

Output Text: Unhappy with the Product

Negative

Figure 6.3 Output for negative review

When the application is given Neutral review as input:

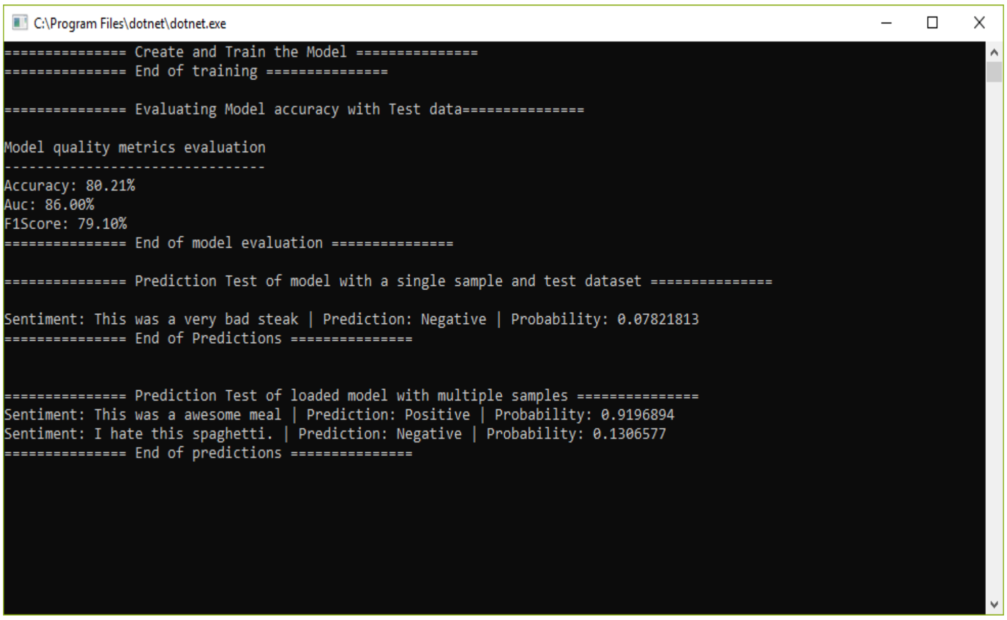


F:\test3.wav

Output Text: its ok ok product

Neutral

Figure.6.3 Output diagram for neutral review



```
============== Create and Train the Model ==============
============== End of training ==============

============== Evaluating Model accuracy with Test data==============

Model quality metrics evaluation
--------------------------------
Accuracy: 80.21%
Auc: 86.00%
F1Score: 79.10%
============== End of model evaluation ==============

============== Prediction Test of model with a single sample and test dataset ==============

Sentiment: This was a very bad steak | Prediction: Negative | Probability: 0.07821813
============== End of Predictions ==============


============== Prediction Test of loaded model with multiple samples ==============
Sentiment: This was a awesome meal | Prediction: Positive | Probability: 0.9196894
Sentiment: I hate this spaghetti. | Prediction: Negative | Probability: 0.1306577
============== End of predictions ==============
```

Figure 6.4 Training model diagram

For the audio classifier, I use in built System (dynamic link library) Speech reference to be used to classifier the features and used in training the model.Table 6.1 shows Comparison between the accuracy obtain from existing hybrid approach and proposed approach and Table 6.2 Confusion Matrix for Novel approach showing correctly match emotion in percentage

Table 6.1 Comparison between the accuracy obtain from existing hybrid approach and proposed approach as shown below:

Experiment	Accuracy
Existing approach using text features[1]	62%
Existing approach using audio features[1]	51%
Proposed approach	80.21%

Table 6.2 Confusion Matrix for Novel approach showing correctly match emotion in percentage

Emotion	Correctly Classified
Positive	84.1%
Neutral	80.1%
Negative	82.4%

Future Scope

As system accepts only speech input there is a limitation on the availability of reviews, therefore it can be extended to multiple input format.

VII. CONCLUSION

In this paper, we have proposed a new hybrid approach to detect the emotions from human utterances. We consider a new technique of combining audio and text features. Our application will depict that the proposed approach entitles better accuracy compared to text or speech mining considered individually. The paper will lead to more realistic human-machine interaction, as it helps to improve the efficiency of emotion recognition of human speech.

ACKNOWLEDGMENT

I wish to express a true sense of gratitude towards my teacher and guide who at every discrete step in the study of this paper, contributed her valuable guidance and help me to solve every problem that arose.

Most likely I would like to express my sincere gratitude towards my family and friends for always being there when I needed them the most.

With all respect and gratitude, I would like to thank all authors listed and not listed in references whose concepts are studied and used by me whenever required.

REFERENCES

[1]. J. Bhaskar, K. Sruthi and P. Nedungadi, Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining, International Conference on Information and Communication Technologies; 2015,pp. 635-643.

[2]. A. Houjeij, L. Hamieh, N. Mehdi and H. Hajj, A Novel Approach for Emotion Classification based on Fusion of Text and Speech, 19th International Conference on Telecommunications; 2012, pp. 5-12.

[3]. R. Hoyo, I. Hupont, F. Lacueva and D. Abad , Hybrid Text Affect Sensing System for Emotional Language Analysis, ACM International Workshop on Affective-Aware Virtual Agents and Social Robots; 2009, pp. 14-16.

[4]. X. Hu, J. Downie and A. Ehmann, Lyric Text Mining in Music Mood Classification, 10th International Symposium on Music Information Retrieval, Kobe; 2009, pp. 411-416.

[5]. M. Ghosh and A. Kar, Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using SentiWordNet 3.0, International Journal of Engineering Research and Technology; 2013, pp. 2- 9.

[6]. T. Vogt, E. Andre and J. Wagner, Automatic Recognition of Emotions from Speech a Review of the Literature and Recommendations for Practical Realisation, Affect and Emotion in Human-Computer Interaction; 2008, pp. 75-91.

[7]. S. Casale, A. Russo and G. Scebba, Speech Emotion Classification using Machine Learning Algorithms, IEEE International Conference on Semantic Computing; 2008, pp. 158-165.

[8]. T. Vogt, E. Andre, and J. Wagner, "Automatic Recognition of Emotions from Speech: a Review of the Literature and Recommendations for Practical Realisation," Affect and Emotion in Human-Computer Interaction, pp.75–91,2008.

[9]. R. del Hoyo, I. Hupont, F. Lacueva, and D. Abad´ıa, "Hybrid Text Affect Sensing System for Emotional Language Analysis," in Proceedings of the ACM International Workshop on Affective-Aware Virtual Agents and Social Robots2009, pp. 1–4.

[10]. E. Vayrynen, J. Toivanen, and T. Seppanen, "Classification of Emotion in Spoken Finnish Using Vowel Length Segments: Increasing Reliability with a Fusion Technique," Speech Communication, 2010 vol. 53, no.3, pp. 269-282.

[11]. X. Hu and J. Downie, "Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio," in Proceedings of the ACM 10th annual joint conference on Digital libraries. 2010, pp. 159–168.

[12]. C. Laurier, J. Grivolla, and P. Herrera, "Multimodal Music Mood Classification Using Audio and Lyrics," Seventh International Conference on Machine Learning and Applications (ICMLA'08) 2008, pp. 688–693.

[13]. T. Kucukyilmaz, B. Cambazoglu, C. Aykanat, and F. Can, "Chat Mining: Predicting User and Message Attributes in Computer- Mediated Communication," Information Processing & Management, vol. 44, no. 4, pp. 1448–1466, 2008.

[14]. H. Binali, V. Potdar, and C. Wu, "A State of the Art Opinion Mining and its Application Domains," in Industrial Technology, 2009. ICIT 2009. IEEE International Conference. 2009, pp. 1–6.

[15]. Various online media like Stack overflow, YouTube, codeproject.