IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

Sentiment Analysis of Speech

Aishwarya Murarka[1], Kajal Shivarkar[2], Sneha[3], Vani Gupta[4],Prof.Lata Sankpal[5]

Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India[1-4]

Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India[5]

Abstract: Communication through voice is one of the main components of affective computing in human-computer interaction. In this type of interaction, properly comprehending the meanings of the words or the linguistic category and recognizing the emotion included in the speech is essential for enhancing the performance. In order to model the emotional state, the speech waves are utilized, which bear signals standing for emotions such as boredom, fear, joy and sadness etc...So we can find different speech signals of each subject. The most significant features that transfer the variations in the tone are classified into pitch and intensity categories. We can use, eleven features, namely, pitch, intensity, the first four formants and their bandwidths and standard deviation, are extracted. The proposed method first digitizes the signal to extract the required properties. According to emotional Prosody studies, the tone of every person's voice can be characterized by its pitch, loudness or intensity, timbre, speech rate and pauses, whose changes convey different information from the speaker to the listener.

Keywords: Speaker recognition, vocal emotion recognition, sentimental analysis, Emotion prediction, Text mining.

I. INTRODUCTION

In a large proportion of these videos, people depict their opinions about products, movies, social issues, political issues, etc. The capability of detecting the sentiment of the speaker in the video can serve two basic functions: (i) it can enhance the retrieval of the particular video in question, thereby, increasing its utility, and (ii) the combined sentiment of a large number of videos on a similar topic can help in establishing the general sentiment. It is important to note that automatic sentiment detection using text is a mature area of research, and significant attention has been given to product reviews, we focus our attention on dual sentiment detection in videos based on audio and text analysis. We focus on videos because the nature of speech in these videos is more natural and spontaneous which makes automatic sentiment processing challenging. In Particular, automatic speech recognition (ASR) of natural audio streams and text spoke in audio is difficult and the resulting transcripts are not very accurate. The difficulty stems from a variety of factors including (i) noisy audio due to non-ideal recording conditions, (ii) foreign accents, (iii) spontaneous speech production, and (iv)diverse range of topics. Our approach towards sentiment extraction uses two main systems, namely, Automatic Speech Recognition (ASR) system and text-based sentiment extraction system. For text based sentiment extraction, we propose a new method that uses POS (Part-Of-Speech) tagging to extract text features and Maximum Entropy modelling to predict the polarity of the sentiments (positive or negative) using the text features. An important feature of our method is the ability to identify the individual contributions of the text features towards sentiment estimation. We evaluate the proposed sentiment estimation on both publically available text databases and videos. On the text datasets, This provides us with the capability of identifying key words/phrases within the video that carry important information. By indexing these key words/phrases, retrieval systems can enhance the ability of users to search for relevant information.

II. LITERATURE SURVEY

From paper*"* A Study of Support Vector Machines for Emotional Speech Recognition" In this paper, efficiency comparison of Support Vector Machines (SVM) and Binary Support Vector Machines (BSVM) techniques in utterance-based emotion recognition is studied. Acoustic features including energy, Mel-Frequency Cepstral coefficients (MFCC), Perceptual Linear Predictive (PLP), Filter Bank (FBANK), pitch, their first and second derivatives are used as frame-based features.[1]

In paper *"*Audio and Text based multimodal sentiment analysis using features extracted from selective regions and deep neural networks" An improved multimodal approach to detect the sentiment of products based on their multi-modality natures (audio and text) is proposed. The basic goal is to classify the input data as either positive or negative sentiment. Learning utterance-level representations for speech emotion and age/gender recognition. Accurately recognizing speaker emotion and age/gender from speech can provide better user experience for many spoken dialogue systems. In this study, we propose to use Deep Neural Networks (DNNs) to encode each utterance into a fixed-length vector by pooling the activations of the last hidden layer over time.[2]

The paper "Towards Real-time Speech Emotion Recognition using Deep Neural Networks" proposes a real-time SER system based on end-to-end deep learning. Namely, a Deep Neural Network (DNN) that recognizes emotions from a

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

one second frame of raw speech spectrograms is presented and investigated. This is achievable due to a deep hierarchical architecture, data augmentation, and sensible regularization. Promising results are reported on two databases which are the ENTERFACE database and the Surrey Audio-Visual Expressed Emotion (SAVEE) database.[4]

From paper "Machine Learning and Sentiment Analysis Approaches for the Analysis of Parliamentary Debates" the author seeks to establish the most appropriate mechanism for conducting sentiment analysis with respect to political debates; Firstly so as to predict their outcome and secondly to support a mechanism to provide for the visualisation of such debates in the context of further analysis. To this end two alternative approaches are considered, a classification-based approach and a lexicon-based approach. In the context of the second approach both generic and domain specific sentiment lexicons are considered. Two techniques to generating domain-specific sentiment lexicons are also proposed: (i) direct generation and (ii) adaptation. The first was founded on the idea of generating a dedicated lexicon directly from labelled source data. The second approach was founded on the idea of using an existing general purpose lexicon and adapting this so that it becomes a specialised lexicon with respect to some domain. The operation of both the generic and domain specific sentiment lexicons are compared with the classification-based approach. The comparison between the potential sentiment mining approaches was conducted by predicting the attitude of individual debaters (speakers) in political debates (using a corpus of labelled political speeches extracted from political debate transcripts taken from the proceedings of the UK House of Commons). The reported comparison indicates that the attitude of speakers can be effectively predicted using sentiment mining. The author then goes on to propose a framework, the Debate Graph Extraction (DGE) framework, for extracting debate graphs from transcripts of political debates. The idea is to represent the structure of a debate as a graph with speakers as nodes and\exchanges" as links. Links between nodes were established according to the exchanges between the speeches. Nodes were labelled according to the \attitude" (sentiment) of the speakers, \positive" or \negative", using one of the three proposed sentiment mining approaches. The attitude of the speakers was then used to label the graph links as being either \supporting" or \opposing". If both speakers had the same attitude (both \positive" or both \negative") the link was labelled as being \supporting"; otherwise the link was labelled as being \opposing". The resulting graphs capture the abstract representation of a debate where two opposing factions exchange arguments on related content's Finally, the author moves to discuss mechanisms whereby debate graphs can be structurally analysed using network mathematics and community detection techniques. To this end the debate graphs were conceptualised as networks in order to conduct appropriate network analysis. The significance was that the network mathematics and community detection processes can draw conclusions about the general properties of debates in parliamentary practice through the exploration of the embedded patterns of connectivity and reactivity between the exchanging nodes (speakers).[3]

In paper" Sentiment extraction from natural audio streams" a system for automatic sentiment detection in natural audio streams such as those found in YouTube is proposed. The proposed technique uses POS (part of speech) tagging and Maximum Entropy modelling (ME) to develop a text-based sentiment detection model. Additionally, we propose attuning technique which dramatically reduces the number of model parameters in ME while retaining classification capability. Finally, using decoded ASR (automatic speech recognition) transcripts and the ME sentiment model, the proposed system is able to estimate the sentiment in the YouTube video. In our experimental evaluation, we obtain encouraging classification accuracy given the challenging nature of the data. Our results show that it is possible to perform sentiment analysis on natural spontaneous speech data despite poor WER (word error rates).[5]

This paper "Techniques and Applications of Emotion
Recognition in Speech" gives a brief overview of the current state of the research in this area with the aim to underline different techniques that are being used for detecting emotional states in vocal expressions. Furthermore, approaches for extracting speech features from speech datasets and machine learning methods with special emphasis on classifiers are analysed. In addition to the mentioned techniques, this paper also gives an outline of the areas where
emotion recognition could be utilised such as healthcare, psychology, cognitive sciences and marketing.[6]

III. IDENTIFY, RESEARCH AND COLLECT IDEA

Political Sentiment Mining Using Classification:-
A] Algorithm

Text based sentiment analysis

The general idea was to use machine learning classifiers trained (learned) using an appropriately labelled training dataset and evaluated using test data. The generated classifiers were then used to predict the attitude of individual speakers participating in an unseen debate. The input is a set of concatenated speeches that make up a single debate and the output is a set of attitude labels one per concatenated speech.

Input:- S = {s1, s2.....sn}
Output:- C = {c1,c2..... cn}

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

S-speeches
C- class labels taken from the set {positive, negative} such that there is a one-to-one correspondence between the elements in S and C.
The process encompasses two phases:
(i) Pre-processing
(ii) Attitude prediction

Pre-processing:-
1. Upper-case alphabetic characters were converted to lower-case letters followed by numeric digit removal.
2. This was followed by a tokenisation Process. The resulting tokens were then indexed to form an initial Bag-Of-Words (BOW = {t1, t2.....tjBOWj}). The next step was to reduce the size of the BOW by removing \stop words".
3. After the completion of stop word removal, each document was represented by some subset of the BOW. Given a specific domain there will also be additional words, other than stop words, that occur frequently. In the case of the House of Commons parliamentary debates words like: \hon.", \house",\minister", \government", \gentleman", \friend" and \member" are all very frequently occurring words. For similar reasons as for stop word removal these domain specific words were also removed. This was done by appending them to the stop-words list.
4. The size of the produced BOW was then further reduced by applying stemming. Stemming is concerned with the process of deriving the \stem" of a given word.
5. On completion of the pre-processing and stemming stages the resulting BOW defines a feature space from which sets of feature vectors can be generated. The feature vector elements hold term weightings. The most widely used mechanism for generating term weightings, and that adopted with respect to the work described in this chapter, is the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme which aims to \balance out the effect of very rare and very frequent" terms in a vocabulary.
6. Thus after the completion of the pre-processing phase the input collection of concatenated speeches were represented using a vector space model such that each speech was described by a feature vector. speech i is represented as a vector

$$Vi = \{wi1, wi2.....wim\}$$

Where, wij is the TF-IDF value for term j in speech i.
7. Once the input data was translated into the feature vector format, whereby the concatenated speeches for each speaker were defined by a subset of words contained in the BOW, classification could be applied to determine each speaker's \attitude" (positive or negative). To this end, a classifier was required. Classifier generation is a supervised machine learning mechanism requires pre-labelled training data. Here we use SVM classifier.

Attitude Prediction:-
Attitude identification using trained classifier

1: INPUT: Set of Vectors V = {v1,v2....vz},
A classifier
2: OUTPUT: Set of Attitudes C = {c1, c2...cz}
where ci $\in$ {positive,negative}
3: C = {}
4: for all vi $\in$ V do
5: ci = Classify(vi) into the fittest class
6: C = C$\cup$ci
7: end for

Audio-based Sentiment Analysis
Audio features like pitch, intensity and loudness are extracted using Open- EAR software and Support Vector Machine (SVM) classifier is built to detect the sentiment . The audio features are automatically extracted from each video clip using OpenEAR software and Hidden Markov Models (HMM) classifier is built to detect the sentiment.

Instead of extracting all the features from the entire input using tools like OpenEAR/ OpenSMILE only specific relevant features like MFCC, prosody and relative prosody are extracted from stressed and normal regions of an input are used in our study.

a) Pre-processing step*:*
The input data in this research is utterances. We have to divide speech signal into frames. Then we compare each frame with phonemes label in the database and find the frame have a silence phoneme label and remove that frame. After that, we merge whole speech frames into utterances again. Silence is considered as an useless data in this research. After getting speech data, all of them are divided into frames. Each frame will be extracted features in below.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

b) Acoustic feature extraction step:

There are two kinds for feature extraction, which are frame-based feature and utterance-based feature. Features such as energy, pitch, Mel-frequency Cepstral Coefficients (MFCC), Perceptual linear predictive (PLP), filter bank (F Bank), and first and second derivatives of all features stated above are extracted as frame based feature. While utterance-based features are calculated the statistical values like maximum, minimum, mean and variance of those frame-based features. In summary, there are four times from converting frame-based feature to utterance-based feature. All experiments in this paper are conducted only with utterance based features on training and testing the classifiers. All of utterance-based features are concatenated together, before calculating the first and second derivatives of them.

Hidden Markov Model (HMM):-

The HMM consist of the first order Markov chain whose states are hidden from the observer therefore the internal behaviour of the model remains hidden. The hidden states of the model capture the temporal structure of the data. Hidden Markov Models are statistical models that describe the sequences of events. HMM is having the advantage that the temporal dynamics of the speech features can be trapped due to the presence of the state transition matrix. During clustering, a speech signal is taken and the probability for each speech signal provided to the model is calculated. An output of the classifier is based on the maximum probability that the model has been generated this signal. For the emotion recognition using HMM, first the database is sort out according to the mode of classification and then the features from input waveform are extracted. These features are then added to database. The transition matrix and emission matrix has been made according to the modes, which generates the random sequence of states and emissions from the model. Final is estimating the state sequence probability by using

Viterbi algorithm.

Support Vector Machines (SVM):-

The support vector machine is a learning algorithm which addresses the general problem of learning to discriminate between positive & negative members of given n-dimensional vectors. The SVM is used for classification & regression purpose.

- The main idea of SVM classification is to a transform the original input set to a high dimensional feature space.
- In Classification, training examples are used to learn a model that can classify the data samples into known classes.
- The Classification process involves following steps:

a. Create training data set.
b. Identify class attribute and classes.
c. Identify useful attributes for classification (Relevance analysis).
d. Learn a model using training examples in Training set.
e. Use the model to classify the unknown data samples.

SVM is a supervised learning process comprising of two steps:

i. Learning (Training): Learn a model using the training data.
ii. Testing: Test the model using unseen test data to assess the model accuracy.

We are proposing sentiment analysis based on video and text. Classifier is used for classification of audio and text.

B] Proposed system

In this architecture user register himself/ herself in our application. For registration he/she should provide personal details. After successful registration user can login to the system. Then system will send the encrypted password to email so that password is prevented from visualization. After successful login, user has the privileges to upload a new video and perform the analysis of the video. A new audio can be recorded which will get stored as a file for analysis. Admin has the privileges to modify the dataset as per requirements.

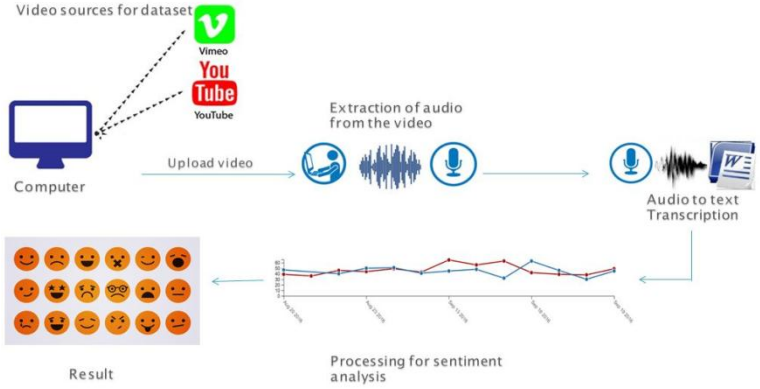

Fig1:- System Architecture

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

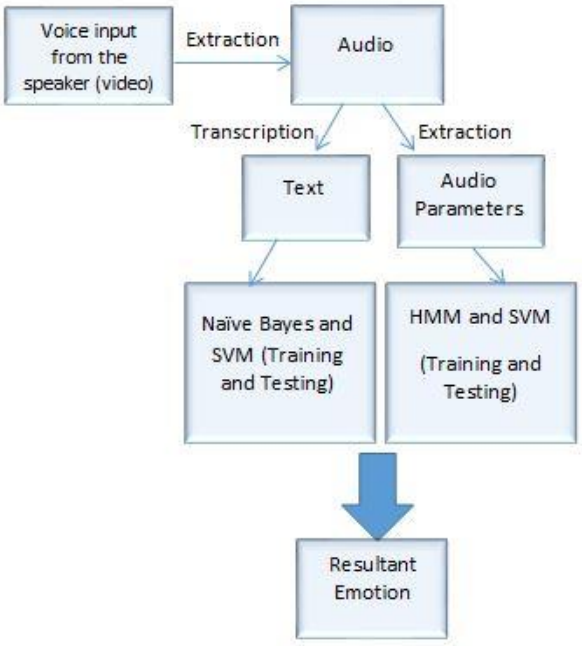

Fig2:- Block Diagram of the Proposed System

IV. SCOPE OF PROJECT

Proposed system uses the dataset consisting of videos based on political speeches. The application of sentiment analysis techniques to predict the attitude of individual debaters. Increasing the spectrum of sentiment classes may provide valuable information, which is not captured efficiently earlier.
Example- Anger, Anxiety, Elation, Confidence etc, instead of Positive Negative and Neutral.

V. CONCLUSION

We believe that multimodality will also help in detecting whether a speaker is expressing his own opinion or merely parroting somebody else's views. In such cases a mere text based approach will fail, as the most important clues will be found in intonation and facial expressions. Hence multimodality can be used in multiple applications in a broader spectrum such as lie detection ,analyzing interviews, interrogations etc. Multimodal Sentiment Analysis is very much an open ended topic. Lots more research needs to be done as evident from the results of the discussed experiment.

V. REFERENCES

[1] Nattapong Kurpukdee , Sawit Kasuriya , Vataya Chunwijitra ,Chai Wutiwiwatchai and Poonlap Lamsrichan ," A Study of Support Vector Machines for Emotional Speech Recognition", 978-1- 5090-4809- 0/17/$31.00 ©2017 IEEE
[2] Harika Abburi," Audio and Text based Multimodal Sentiment Analysis using Features Extracted from Selective Regions and Deep Neural Networks", International Institute of Information Technology Hyderabad - 500 032, INDIA June 2017
[3] Zaher Ibrahim Saleh Salah," Machine Learning and Sentiment Analysis Approaches for the Analysis of Parliamentary Debates", May 2014
[4] "Towards Real time speech emotion recognition using deep neural network"2017
[5] Lakshmish Kaushik, Abhijeet Sangwan, John H. L. Hansen," SENTIMENT EXTRACTION FROM NATURAL AUDIO STREAMS", 978-1-4799-0356- 6/13/$31.00 ©2013 IEEE
[6] S. Lugović, I. Dunđer and M. Horvat,"Techniques and Applications of Emotion Recognition in Speech", MIPRO 2016, May 30 - June 3, 2016, Opatija, Croatia