AUTOMATIC SENTIMENT DETECTION IN TEXT AND NATURALISTIC AUDIO

1. AMRUTHA K V, 2. HEMA KRISHNAN

1. .PG STUDENT, 2. ASSISTANT PROFESSOR CUM RESEARCH SCHOLAR

DEPARTMENT OF COMPUTER SCIENCE,

FISAT, ANGAMALY, ERNAMKULAM, INDIA

Abstract : ***Audio sentiment analysis utilizing automatic speech recognition is a rising research area where opinion or sentiment shown by a speaker is distinguished from normal sound. It is generally underexplored when contrasted with text based sentiment detection. Extracting speaker sentiment from nat- ural audio sources is a challenging problem. Conventional techniques for sentiment extraction generally use transcripts from a speech recognition system, and process the transcript utilizing text based sentiment classifiers. This paper proposes a text based sentiment classifier which decides the most valuable and discriminative sentiment bearing keyword terms, utilized as a term list for KWS. To get a conservative yet discriminative sentiment term list, iterative feature optimization for maxi- mum entropy sentiment model is proposed to reduce model complexity while keeping up effective classification accuracy. A new hybrid ME-KWS joint scoring methodology is created to display both text and audio based parameters in a single integrated formulation. After this, a comparison study was conducted with SVM classifier which indicates moderately poor accuracy. Test results demonstrate that the proposed KWS based framework fundamentally beats the customary ASR architecture in detecting sentiment for challenging practical undertakings.***

IndexTerms **– Sentiment Detection, Keyword Spotting(KWS),Maximum Entropy(ME),SVM Classifier.**

I. INTRODUCTION

Text based sentiment detection is a built up field in Natural Language Processing (NLP). Sentiment analysis /opinion mining, analyses peoples opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. There is an enormous measure of obstinate information in the web-based social networking and on the Web as Twitter, Facebook, message sheets, online journals, and client discussions. Amazon, Yahoo, Google and different other customized sites are a critical asset for acquiring opinion concerning results of any sort. Numerous purchasers structure their choice to purchase an item subject to input from online surveys. This data not only helps ordinary individuals decide, yet additionally gives markers to organizations about the gathering of an item, or a political setting, to comprehend the state of mind of individuals with respect to a continuous social/cultural/political/economic issue. Normally, a given content is ordered to show positive, negative or neutral sentiment.

Text based reviews form only one of the many ways people can express their sentiment/opinion about products or social issues. Audio/Video is additionally a noticeable technique to express sentiments. There are numerous audio platforms on the Web where people express their feelings. Likewise, the audio mode is all the more dominant than text for some circumstances since they give more extravagant signs of the speaker in regards to their conclusions. Detecting sentiment in audio is still an unexplored area.

Speech based sentiment extraction is an emerging and challenging field. A hybrid system is developed which utilizes a robust Automatic Speech Recognition (ASR) system in tandem with NLP based sentiment analysis techniques to detect sentiment of audio streams. Not at all like text based sources, audio sources have a high level of fluctuation both in terms of expressing opinion just as the method of expression of the opinion. There are a scope of difficulties for sentiment extraction in highly natural speech sources including:

- Domain and Vocabulary : The speaker can express opinions about any topic, (e.g., products, movies, politics, social issues, games, etc.) Hence the ASR system should be efficient to handle a wide range of domains and vocabulary. The language model should be comprehensive.
- Speaker Variability and Speaker Accents : ASR system should be robust to speaker variability which includes a wide range of English accents from all over the world.
- Noisy Audio and Channels : Inconsistent recording equipment and different mode/distance of recording, inconsistent acoustic and background environment conditions make the sentiment detection problem challenging. Also, background music/talk, intentional music mixing, reverberation issues make the problem harder.
- Natural and Spontaneous : Detecting audio sentiment in natural and spontaneous speaker settings and various speaker interactive scenarios (i.e., 1-way, 2-way, public speech etc.) is challenging.

Given the touchy increment of online recordings on item reviews, un-boxing, politics, sports, culture, and so on sites such as YouTube.com, Vimeo, News broadcasting, Daily Motion, Twitch and Vine, automatic audio sentiment detection technology would be helpful in gathering and outlining information for clients. Sentiment analysis can be classified mainly into the following four categories:

- Document-level sentiment analysis : The sentiment is generated on the overall document/review level. This is a global analysis.
- Sentence-level sentiment analysis : This gives a microlevel sentiment assessment for every sentence. This is effectively a local analysis.
- Aspect-based sentiment analysis : This gives a sentiment variation in both a local and global level. Aspect-based sentiment analysis focuses on the recognition of all sentiment expressions within a given document. It is possible to have varying local sentiment with one overall document assessed value.
- Comparative sentiment analysis : Extracting opinion in reviews where a product is compared with other product.

This paper contains the following details. A literature survey is conducted in section II. Next, the proposed methodology is explained in section III which includes explanation of each phase in detail. Conclusion and Future work of the proposed method are listed in section IV.

II. RELATED WORKS

In [1], they proposed a system for Automatic sentiment detection in natural audio streams such as those found in YouTube. The proposed technique uses POS(Part of Speech)tagging and Maximum Entropy modeling (ME) to develop a text based sentiment detection model. Additionally, they propose a tuning technique which dramatically reduces the number of model parameters in ME while retaining classification capability. Finally, using decoded ASR(Automatic Speech Recognition) transcripts and ME sentiment model, the proposed system is able to estimate the sentiment in the YouTube video.

They use the mentioned ASR system to obtain text from YouTube video data. It is useful to note that I-best hypothesis, lattices and nbest-lists can be used in the proposed system. Once the decoded text for the video is obtained, they use the POS tagger based feature extraction technique to identify useful sentiment features. Using these sentiment features, the ME-based sentiment model is used to estimate the sentiment polarity. The final output of our system are the probabilities of positive and negative sentiment (and they sum to unity).

ME model does not use Noun based text features. The motivation behind removing noun features was to prepare a more domain independent sentiment model for YouTube videos. It automatically extract the most unambiguous text features using the proposed ME model with good accuracy.

Lakshmish Kaushik, Abhijeet Sangwan, John H L. Hansen [2] present a system that focus on Automatic Sentiment Extraction from YouTube Videos. The extraction of speaker sentiment from natural audio streams such as YouTube is challenging. In this study, they build upon their previous work, where they had proposed a system for detecting sentiment in YouTube videos. Particularly, they proposed several enhancements including:

- better text-based sentiment model due to training on larger and more diverse dataset
- an iterative scheme to reduce sentiment model complexity with minimal impact on performance accuracy
- better speech recognition due to superior acoustic modeling and focused (domain dependent) vocabulary/language models
- a larger evaluation dataset.

Collectively, their enhancements provide an absolute 10percent improvement over their previous system in terms of sentiment detection accuracy. Additionally, they also present analysis that helps understand the impact of WER (word error rate) on sentiment detection accuracy. Finally, they investigate the relative importance of different Parts-of-Speech (POS) tag features towards sentiment detection. In this study, they have used significantly more data from diverse sources for training their text-based ME sentiment models. Next, they have also proposed a new method that iteratively prunes ambiguous features from the ME based sentiment model. Additionally, this method allows us to continue to increase our training dataset while managing model complexity. Next, they have used a more powerful KALDI based speech recognition engine that uses SAT (speaker adaptive training) acoustic models with a bigger (and more application focused) language model.

This system iteratively prunes ambiguous features and provides a smaller vocabulory. These are the main advantages of this system. It will not automatically extract the demographic information about the speaker of the sentiment and also it will not automatically extract the object at which the sentiment is directed. So this is not an optimal system.

In paper [3], they designed a framework for Automatic Audio Sentiment Extraction Using Keyword Spotting. Most existing methods for audio sentiment analysis use automatic speech recognition to convert speech to text, and feed the textual input to text-based sentiment classifiers. In this study, a single keyword spotting system (KWS) is developed for sentiment detection. A text-based sentiment classifier is utilized to automatically determine the most powerful sentiment-bearing terms, which is then used as the term list for KWS. In order to obtain a compact yet powerful term list, a new method is proposed to reduce text-based sentiment classifier model complexity while maintaining good classification accuracy. Finally, the term list information is utilized to build a more focused language model for the speech recognition system. The result is a single integrated solution which is focused on vocabulary that directly impacts classification. The proposed solution is evaluated on videos from YouTube.com and UT-Opinion corpus. Sentiment detection accuracy depends on being able to reliably detect a very focused vocabulary in the spoken comments. Therefore, Keyword Spotting (KWS) technology seems to be better suited for sentiment detection, as opposed to full-transcript ASR. The language model for the speech recognizer is built offline by using a mixture of sentiment text data and conversational telephony transcripts. Sentiment text data is also used to generate the keyword term list by applying the proposed iterative pruning method. The proposed sentiment detection system uses the term list to search for sentiment bearing terms in the audio. This model is better suited for sentiment detection, as opposed to full-transcript ASR and for high-level semantic classification tasks with good accuracy. It could not focus on using pure speech features to augment lexical information drawn for speech recognition to do speech sentiment detection.

In [4], introduced a system for Sentiment Retrieval on Web Reviews using Spontaneous Natural Speech. This paper addresses the problem of document retrieval based on sentiment polarity criteria. A query based on natural spontaneous speech, expressing an opinion about a certain topic, is used to search a repository of documents containing favorable or unfavorable opinions. The goal is to retrieve documents whose opinions more closely resemble the one in the query. A semantic system based on speech transcripts is augmented with information from full-length text articles. This paper makes three important contributions. They introduced a framework for polarity analysis of sentiments that can accommodate combinations of different modalities capable of dealing with the absence of any modality. It is possible to improve average precision on speech transcriptions sentiment retrieval by means of regularization and demonstrate the robustness of their approach by training regularizers on one dataset, while performing sentiment retrieval experiments, with substantial gains, on another dataset.

In this paper, they proposed a feature regularizer suitable for sentiment polarity analysis through modal expansion. A set of reviews for small electronic appliances was collected, containing a small video clip and a full text article on each product. A linear operator is then learned to minimize the average similarity of data across the two modalities. This acts as a feature regularizer for samples belonging to the noisier of the modalities. The main advantage of their system is to improve sentiment retrieval accuracy even in the absence of one modality.

In [5], presented a framework for Automatic Sentiment Analysis from Opinion of Thais Speech Audio. This paper aims to develop a method of sentiment analysis for Thais customers to identify the different notions into two opinions (positive or negative) to consume the products. These opinions are represented by text that is derived from the Thais speech audio content in social media especially video reviews about beauty product. Then, this work implements the model by the Naive Bayes text classification. The results could be demonstrated that the method can provide more effectiveness and satisfactory accuracy for automatic sentiment analysis. Automatic sentiment analysis was divided into four main parts : sentiment modeling, audio extraction, speech to text and word segmentation, automatic sentiment analysis engine.

In [6], they proposed the work to detect hate speech in the Indonesian language. As far as they know, the research on this subject is still very rare. The only research they found has created a dataset for hate speech against religion, but the quality of this dataset is inadequate. Their research aimed to create a new dataset that covers hate speech in general, including hatred for religion, race, ethnicity, and gender. In addition, they also conducted a preliminary study using machine learning approach. Machine learning so far is the most frequently used approach in classifying text. They compared the performance of several features and machine learning algorithms for hate speech detection. Features that extracted were word n-gram with n=l and n=2, character n-gram with n=3 and n=4, and negative sentiment. The classification was performed using Naive Bayes, Support Vector Machine, Bayesian Logistic Regression, and Random Forest Decision Tree.

Another objective of their research is to compare features and machine learning algorithms to find out which combination of features and algorithm that have the best performance. Their methods consist of three stages: Pre-processing, Feature Extraction, Classification and Evaluation. They conducted three scenarios of experiments. First, they compared the performance of every combination of features and algorithms. Second, they compared the performance of word n-gram vs. character n-gram. Third, they compared the performance of each algorithm when using all the five features altogether.

In Sentiment Analysis on Speaker Specific Speech Data [7], they proposed a method for analysing the sentiment of speaker data. Sentiment analysis has evolved over past few decades, most of the work in it revolved around textual sentiment analysis with text mining techniques. But audio sentiment analysis is still in a nascent stage in the research community. In this proposed research, they perform sentiment analysis on speaker discriminated speech transcripts to detect the emotions of the individual speakers involved in the conversation. They analyzed different techniques to perform speaker discrimination and sentiment analysis to find efficient algorithms to perform this task.

In this paper, they proposed a model for sentiment analysis that utilizes features extracted from the speech signal to detect the emotions of the speakers involved in the conversation. The process involves four steps : Pre-processing, Speech Recognition System, Speaker Recognition System, Sentiment Analysis System. This system is highly scalable with good accuracy. The main drawback of this system was it can not handle conversation between two people talk simultaneously.

In [8], they introduced a system for Human Speech Sentiments Recognition, A Data Mining Approach for Categorization of Speech. This paper proposes gender dependent emotion recognition system. They have collected a huge database that consists of emotional speech generated by males and females under different age groups. Based upon the speech produced by different people, each entity speech is then classified into different categories such as happy, sad, anger, bored and neutral. Various features extraction techniques is applied to extract the features from speech. The classification techniques such as support vector machine, linear discriminant analysis classifier, K-nearest neighbors and naive bayes classifier are then applied on the extracted features to classify the emotions. A gender classification system is adopted based on Gaussian Mixture Model for classifying the gender. Results show that the accuracy of emotion classification system is above 99percent if gender is known. For emotion recognition, we have mentioned two subsystems: Gender Recognition selection/reduction algorithms and test designated classifiers. For Gender recognition we have used Gaussian Mixture Model classifier. Speech signals are used for the extraction of features. These numerous extracted parameters are passed through classifiers so as to discover the emotion class. Various classifiers are used for detecting emotions from speech.

[9] provide an alternative solution to support Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition. Speech recognition and artificial intelligence powers the automatic speech recognition systems, these systems can be applied in the call center environments. This paper explains the use of sentiment analysis to identify if the customer are satisfied the ASR systems performance. This paper presents approaches and techniques for how sentiment analysis can be used in call centre environments to recognize user emotions.

Improved Multimodal Sentiment Detection Using Stressed Regions of Audio [10] presented an improved approach to detect the sentiment of an online spoken reviews based on its multi-modality natures (audio and text). To extract the sentiment from audio, Mel Frequency Cepstral Coefficients (MFCC) features are extracted at stressed significant regions which are detected based on the strength of excitation. Gaussian Mixture Models (GMM) classifier is employed to develop a sentiment model using these features. From results, it is observed that MFCC features extracted at stressed significance regions perform better than the features extracted from the whole audio input. Further from the transcript of the audio input, textual features are computed by Doc2vec vectors. Support Vector Machine (SVM) classifier is used to develop a sentiment model using these textual features. From experimental results it is observed that combining both the audio and text features results in improvement in the performance for detecting the sentiment of a review.

III. PROPOSED SYSTEM

In the proposed audio based sentiment detection model, there are three core subsystems. First one is the offline text based sentiment model generation, the second is the ASR based sentient detection system forming our baseline system, and finally the third is a proposed system using audio Keyword Spotting (KWS) approach. After this , a comparative study was conducted with SVM classifier and it shows poor accuracy when compared to HMM classifier.

Generic methods for sentiment extraction generally use transcripts from a speech recognition system, and process the transcript using text-based sentiment classifiers. A new architecture using keyword spotting (KWS) is proposed for sentiment detection. In the new architecture, a text-based sentiment classifier called Hidden Markov Model(HMM) is used to automatically determine the most useful and discriminative sentiment-bearing keyword terms, which are then used as a term list for KWS. Inorder to obtain a compact yet discriminative sentiment term list, iterative feature optimization for maximum entropy sentiment model is proposed to reduce model complexity while maintaining effective classification accuracy. After this, a comparison study was conducted with another text-based sentiment classifier called SVM and it shows moderately poor accuracy.

The baseline of the proposed sentiment detection system is shown in Fig 1. This system consists six parts : Data collection, Pre-processing, Feature Extraction, Training, Testing, and Prediction of sentiment. Firstly, text and audio dataset of different emotions were collected. After that pre-processing was done on text dataset using NLTK and on audio dataset using array conversion. In the case of text, features are extracted using Bag of Words(BoW) and these words are then converted to vector form. Similarly, features are extracted from audio using Mel Frequency Cepstrum Coefficient(MFCC). After feature extraction, performed training on text using Maximum Entropy classifier and on audio using HMM and SVM classifier. Testing on text was done by manually entering a sentence and it is on audio was done by providing a path to an audio. After detecting the emotion, the system can predict the sentiment(positive or negative).

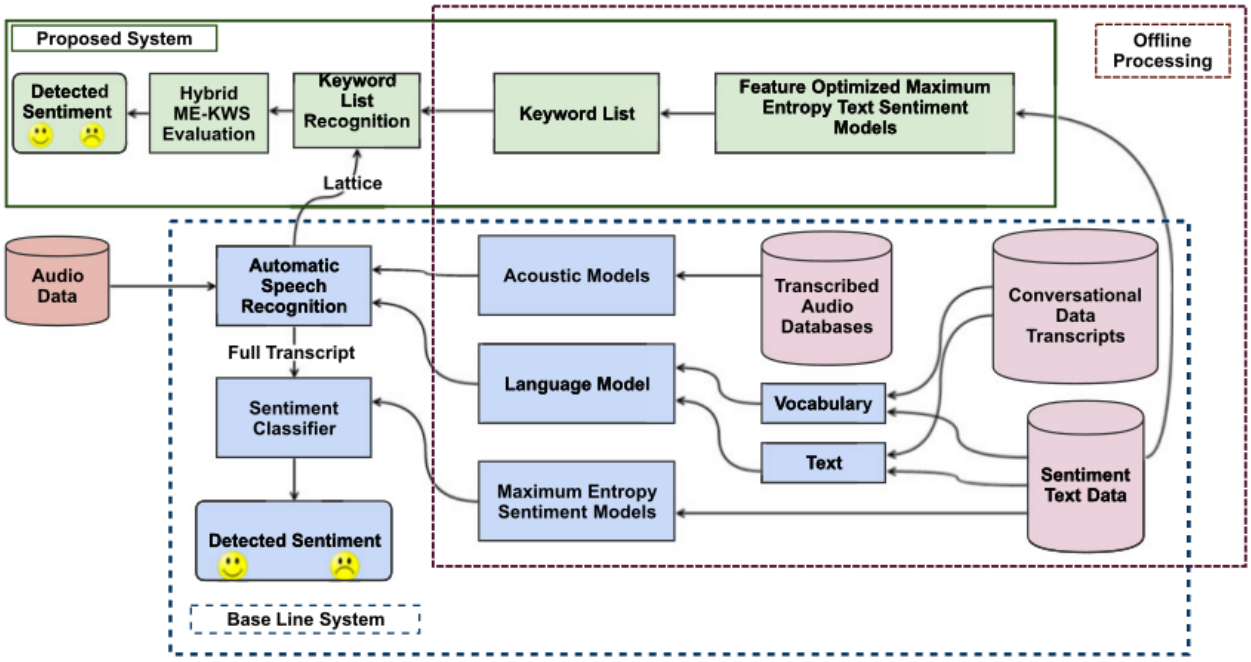

Figure 1. Baseline of proposed method

3.1 DATA COLLECTION

Text dataset containing different emotions like happy, sad, fear, disgust, anger, shame, neutral and audio dataset containing emotions like happy, sad, fear, anger, disgust and neutral were collected.

3.2 PREPROCESSING

3.2.1 Text preprocessing

- converting all letters to lower or upper case
- converting numbers into words or removing numbers
- removing punctuations, accent marks and other diacritics
- removing white spaces
- expanding abbreviations
- removing stop words, sparse terms, and particular words using NLTK
- text canonicalization

3.2.2 Audio preprocessing

In audio preprocessing, The following steps are considered:

- Covert .mp4 audio file to .wav file.
- First step will be to read the audio file from given path
- Directory iterator that handle .wav files by converting them to numpy arrays
- calculate the spectrogram of the audio file**.**

3.3 FEATURE EXTRACTION

The feature extraction part of the framework is the base part of the system. Feature set must choose in such a way that, it must be fast and efficient to reach a conclusion after performing data analysis on the features.

3.3.1 Feature Extraction(Text)

BoW is used for feature extraction. It is a way of extracting features from text for use in modeling, and this approach is very simple and flexible. A bag-of- words is a representation of text that describes the occurrence of words within a document. It involves two things: A vocabulary of known words and a measure of the presence of known words. In this approach histogram of the words within the text are considered by taking word count as a feature. In text feature extraction, the following steps are considered:

- Collect data
- Design the Vocabulary
- Create Document Vectors

3.3.2 Feature Extraction(Audio)

In audio feature extraction, the following steps are considered:

- First step will be to read the .wav audio file from given path
- Take audio input from one channel only
- Perform downsample operation
- Compute MFCC using librosa library
- MFCC vectors might vary in size for different audio input. Prepare a fixed size vector for all of the audio files.
- To overcome this problem , need to pad the output vectors with constant value (the one I used here is 0).

Since computing MFCC is time consuming, then do it only once. And for later times just load it from the saved files.

3.4 TRAINING

3.4.1 Text

For training text data, Maximum Entropy Classifier is used. Maximum Entropy (ME) is a multinomial logistic regression method that predicts the probabilities of different possible outcomes of a categorically distributed dependent variable, given a set of independent variables. Maximum entropy based sentiment system is an effective approach for discriminative learning of features. This can be used in huge database scenarios to develop probabilistic models that can detect the sentiment effectively. Consider sentiment classification as a two-class problem, (i.e., Positive vs. Negative classification), and the Maximum Entropy based approach is used to develop the sentiment classifier. Iterative Maximum Entropy Optimization is used to learn sentiment classification models. After this, a .pickle file is created which is normally considered as the model for the proposed system.

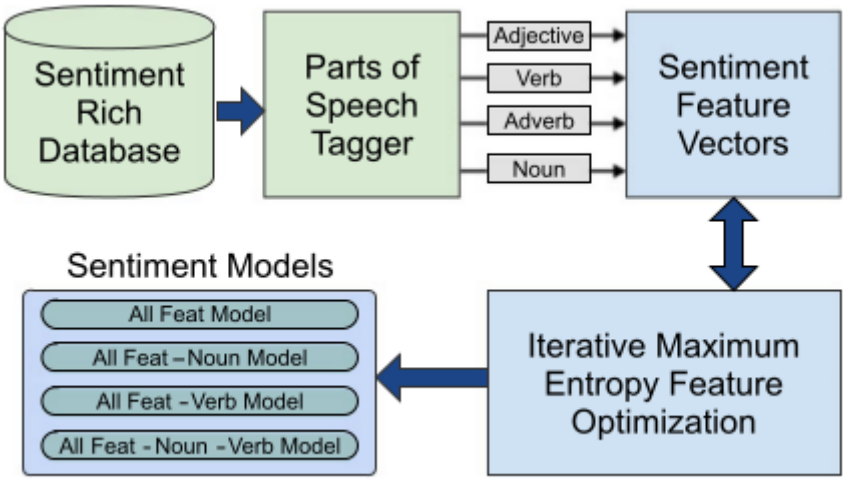

Figure 2. Steps for generating sentiment models

Figure 2 shows the Steps for generating sentiment models using the proposed Iterative Maximum Entropy Optimization algorithm. Using a sentiment rich text database, parts-of-speech tags (i.e., Adjective, Verb, Adverb, Nouns combinations) are used to extract sentiment features. Iterative Maximum Entropy Optimization is used to learn sentiment classification models.

3.4.1 Audio

A simple way to perform sentiment detection for audio data is to process ASR output transcripts through a text-based sentiment detection system (such as the ME method described in previous sections). In this study, this system constitutes the baseline for audio based sentiment detection experiments. The acoustic models use Mel frequency cepstral coefficients (MFCCs) for feature extraction and Hidden Markov Model(HMM) for training. For comparison study, SVM is also used for training which shows less accuracy compared to HMM. Steps involved in the creation of HMM model are listed below:

- Create models with passed parameters
- Trains an HMM for each class and it follows sklearn style.
- Predicts labels on an unseen test set. Label from HMM with highest score is chosen.
  Similarly, SVM is used for training of audio data. After this , a model named testfile.sav is obtained. Using this, accuracy, precision, recall, F1-score and support are computed.

3.5 TESTING AND PREDICTION

3.5.1 Text

The steps involved in testing are listed below:

- Read the sentence (enter the text manually)
- Reads input dataset and number of reviews in it
- Perform preprocessing
- When compared the test data with the existing data, system will return features of most sentiment bearing terms
- Using ME classifier, emotion of the test data can be predicted
- From the returned emotion, system can predict whether the emotion belongs to positive or negative.

3.5.2 Audio

- Provide the path of audio file
- Perform preprocessing
- Corresponding features are extracted using MFCC
- Load the created model
- Using HMM classifier, emotion of the test data can be predicted.
- From the predicted emotion, system returns positive or negative.

IV. CONCLUSION

A new architecture using keyword spotting (KWS) is proposed for sentiment detection. In the new architecture, a text-based sentiment classifier is utilized to automatically determine the most useful and discriminative sentiment-bearing keyword terms, which are then used as a term list for KWS. By focussing on the terms that impact decision and ignoring non-sentiment bearing words/phrases, the overall system is more immune to speech recognition errors. Additionally, a new method to create the sentiment bearing keyword list for KWS has also been proposed. The method uses an iterative methodology to automatically extract sentiment bearing keywords from text. A new method for sentiment scoring that combines keyword spotting likelihood (or confidence) into Maximum Entropy likelihood computation has also been proposed. A comparison study of this result with SVM classifier is shown in this project. Accuracy of SVM classifier is very poor compared to the accuracy of HMM.

REFERENCES

[1] Lakshmish Kaushik, Abhijeet Sangwan, John H L. Hansen, Sentiment Extraction from Natural Audio Streams, IEEE, 2013.

[2] Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen, Automatic Sentiment Extraction from YouTube Videos, IEEE, 2013.

[3] Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen, Automatic Audio Sentiment Extraction Using Keyword Spotting, IEEE, 2015.

[4] Jose Costa Pereira, Jordi Luque , Xavier Anguera, Sentiment Retrieval on Web Reviews using Natural Spontaneous Speech, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[5] Preedawon Kadmateekarun, Sumitra Nuanmeesri, Automatic Sentiment Analysis from Opinion of Thais Speech Audio, IEEE International Conference on Science and Technology, 2015.

[6] Ika A lfina, Rio M ulia, M ohamad Ivan Fanany, and Yudo Ekanata, Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study, ICACSIS, IEEE, 2017.

[7] Maghilnan S, Rajesh Kumar M , Sentiment Analysis on Speaker Specific Speech Data, International Conference on Intelligent Computing and Control, 2017.

[8] Ritika Gupta, Gaurav Aggarwal, Human Speech Sentiments Recognition: A Data Mining Approach for Categorization of Speech, IEEE, 2016.

[9] Rohit Raj Sehgal, Shubham Agarwal, Gaurav Raj, Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition, IEEE International Conference on Advances in Computing and Communication Engineering, 2018.

[10] Harika Abburi, Manish Shrivastava and Suryakanth V Gangashetty , Improved Multimodal Sentiment Detection Using Stressed Regions of Audio, IEEE , 2016.