A Framework for Combining Acoustic and Textual Features in Sentiment Analysis

Sureshkumar Govindaraj
Department of Computer Science, Pondicherry University, India

Abstract—Sentiment analysis has been used in widespread applications which helps the customers to analyze product qualities and helps the business people in customer retention, product promotions and customer behavior analysis. Traditionally textual features has been used for sentiment analysis. But, textual features alone cannot capture the deep emotions of customers particularly when customer’s response is in the form of voice. In this paper, a method has been proposed for combining acoustic features and textual features for performing sentiment analysis on customers’ voice response. The proposed method, extracts potential acoustic features that represent various emotions of customers speaks about products. These acoustic emotion features are combined with lexical sentiment features and used in supervised classifier to predict the sentiments of customers. The experiments are conducted with voice clips generated by replicating the product review text documents crawled from amazon.com. The results obtained by well performing supervised classifiers with different combinations of sentiment features are analyzed.

Keywords—Sentiment Analysis, Acoustic Features, Textual Features, Sentiment Classification.

I. INTRODUCTION

Automatic sentiment analysis of voice data can support the business community to understand the true emotions of their customers, which will help them to pay special attention according to customers’ satisfaction levels. The existing researches on sentiment analysis failed to utilize voice emotion features for sentiment analysis. The proposed approach to perform sentiment analysis investigates the effect of acoustic features when it is combined with lexicon based sentiment features. Audio features such as voice intensity, loudness, fundamental frequency (F0) and Mel-Frequency Cepstral Coefficients (MFCC) introduced by [1] has been used in this paper. The audio data is transcribed into text using automatic speech recognition [2], and then SentiWordNet [3] tool has been used for extracting word features such as positive, negative and objectivity scores. The popular classifiers Support Vector Machine is used for the automatic prediction of customer sentiments about products. The sentiments categorized as positive, neutral, and negative.

The acoustic emotion features are used to enhance the sentiment polarity indications obtained from text-based sentiment features by adopting customers’ mood or emotions.

The experiments are conducted with voice clips generated by replicating the product review text documents crawled from amazon.com. The results obtained by well performing supervised classifiers with different combinations of sentiment features are analyzed.

II. RELATED WORK

The most of the existing methods proposed for performing sentiment analysis deals with textual features. There are some extended methods proposed to utilize adverbs and adjective features that enhance the sentiment polarity of sentiment words [4, 5, 6]. The bag-of-word features such as adverbs and adjectives proven to be suitable to use with rule based classifiers [7]. The sentiment analysis methods used in many applications, such as text to speech synthesis [8], opinion mining in on-line forums and electronic news media [9] [10], question answering [11], text summarization [12], and citation analysis [13].

Emotional signals extracted from audio data can be used as a cue for efficient sentiment mining process. Souraya Ezzat [14] proposed a method to predict speakers’ emotion using acoustic features and used multiple classifiers such as Decision Tree, Naïve Bayesian, SVM and K-Nearest for validation. Dragut [15] proposed a method to measure the effect of adverbs on strengthening sentiment scores. The large-scale and opinion-rich data provided by social media applications used to analyze the public sentiments [16]. Marcela Charfuelan and Marc Schröder [17] investigated the possible correlations between sentiment scores obtained for sentences and acoustic features extracted from the same sentences. They found that the calculated acoustic emotional scores more correlated with average energy and mean fundamental frequency (F0). Moghaddam and Ester proposed a method called Opinion Digger [18] that used Part-Of-Speech (POS) tags as features used in an unsupervised machine learning to determine set of aspects.
There are some other acoustic attributes used as the cue to recognize voice emotions are Mel-frequency cepstral coefficients (MFCC), Mel-based speech signal power coefficients, formants and temporal features such as speech rate and pausing [19]. In several methods, prosody features are also used for speech-based emotion recognition [20, 21].

From the survey it is observed that, emotion based sentiment analysis methods in general, aiming at recognizing the mood and sentiment of the people using both conceptual semantics (sentiment words) and the mental as well as physical-state of the people [22].

III. PROPOSED METHOD

The method proposed in this paper for automatic analysis of sentiments, utilizes emoticons features (EF) extracted from product reviews presence in the form of voice data and sentiment features (SF) extracted form of textual transcription of the same audio clips. The objective of this proposal is to identify the opinion of a customer about a product in a more accurate level. The opinion is classified into three types, which reflect the sentiment polarity and emotional status of the customers. The architecture of the proposed method is depicted in Fig. 1. The proposed model takes voice data as input and delivers any one sentiment (opinion) from the three categories; positive, neutral, and negative.

There are 23 features used in this paper are widely used in many researches on emotion recognition. The text feature is extracted by converting the voice into speech text using a speech recognition toolkit. Not all the words presences in speech text are expected to carry sentiment indications. Hence, a lexical parser has been used for extracting sentiment carrying words that is considered as potential sentiment features. The sentiWordNet [3] tool used for calculating sentiment scores. These voice and text features are validated using well performing classifier model to identify the suitable feature set for the prediction of sentiments from voice data.

In this paper, SVM classification model used to analyze the effectiveness of emotion features and sentiment features for sentiment analysis. The proposed method to automatically extract sentiment word features from textual product review document uses the Stanford lexical parser [23] for generating a parse tree for each individual sentence. Then, a binary decision tree-based rule engine applies set of hand-coded rules on the lexical parsing patterns. The outcome of the rule engine is a set of triples consists of three components: aspect, which represents the attributes of product; predicate which links an emotion towards the aspect; and target sentiment words, which are considered as the sentiment carrying words. The set of such triples are used to calculate sentiment scores for training the classifier model, which can be used to predict the true sentiment of the customers on a specific product.

A. Sentiment Features (SF)

The lexical tool SentiWordNet [3] is used for automatically extracting sentiment features from text. SentiWordNet consists of annotated synsets of WordNet according to the notions of “positivity”, “negativity”, and “neutrality”. Each synset is associated to three numerical scores Pos(s), Neg(s), and Obj(s) which indicate the positive, negative, and objective nature of the terms contained in the synset. The objectivity score of a word is calculated according to (1).

\[ \text{ObjScore} = 1 - (\text{PosScore} + \text{NegScore}) \]  

The sentiment feature vector is constructed using the individual sentiment scores calculated for all sentiment carrying words generated during feature extraction process. Also, the mean sentiment score of all sentiment words included for each feature. An un-weighted mean sentiment score obtained from SentiWordNet has been used for all the feature words.
B. Emotion Features (EF)

The open source natural language processing software OpenEAR [1] has been used in this paper for extracting acoustic features from audio clips. The most important voice signal measures such as prosody, energy, voicing, spectrum, and cepstral features are being used as the acoustic features. Prosody features are extracted from frequency and amplitude of the audio signal that represent the intensity, loudness, and pitch. Energy features describe the human loudness perception, and the presence of voice is identified using energy feature. Cepstral features represent changes or periodicity in the spectrum features that is measured by Mel-frequency cepstral coefficients calculated based on the Fourier transform of a speech frame. These acoustic features represent the emotional state of the speaker.

Thus, the following emotion feature scores are calculated to construct the emotion feature vector.

- **Pitch**: Minimum, maximum, mean, standard deviation, absolute value, quantile, ratio between voiced and unvoiced frames.
- **Duration**: Time ($t^E$) and Height ($h^E$).
- **Intensity**: Minimum, maximum, mean, standard deviation, quantile.
- **Formant**: First formant, second formant, third formant, fourth formant, fifth formant, second formant / first formant, third formant / first formant.
- **Rhythm**: Speaking rate.

IV. DATA AND EXPERIMENTS

The dataset used in this paper consists of 120 audio clips recorded in a controlled environment with 10 trained actors. The textual product reviews selected from Amazon product review collection in five different scenarios simulating real customer emotion during a conversation with call-centre agent are selected. These scenarios reflect the customer opinion on a product in three category. The product reviews cover five different products such as mobile phones, mp3 players, digital cameras, shoes and watches. The audio clips are converted again from speech to text for extracting lexical sentiment features. The audio clips used for training the classifiers are pre-labeled depending on the context of the speech, there are 40 instances are used in each category positive, neutral and negative.

For the purpose of feature evaluation, a separate training and test set has been prepared for text sentiment features (SF), voice emotion features (EF) and voice emotion features combined with sentiment features (EF+SF). Each data set has been selected to find the effect of emotion features on intensifying the sentiment classification process. The data sets were investigated using the well performing SVM classifier used in natural language processing researches [24, 25]. The LibSVM [26] multi-class SVM model has been used for assigning three different class labels to the product review data. Multi-class SVM (n-SVM) defines a hyper plane by calculating distance between the finite numbers of feature values. The n-SVM model is used with a RBF kernel with a bias and gamma of 1.0 as Kernel function parameters.

V. RESULTS

The individual data set has been individually evaluated using same classifier model. For the purpose of performance comparison, the test data set are manually evaluated using Gold Standard evaluation strategy. Then, the confusion matrix tables are plotted by comparing classifier outcome with human validation. From the obtained confusion tables the standard performance measures such as the accuracy, precision, and recall scores are calculated for the data sets. The calculated precision, recall and accuracy scores SF, EF and SF+EF feature sets are visualized in Table 1. When the SVM classifier model is tested with lexical sentiment feature (SF) alone, achieved a highest accuracy of 80% for ‘positive’ category and the mean accuracy obtained for all the three categories is 72%. Highest accuracy of 73 % is achieved for ‘positive’ category with only acoustic emotion features (EF), and the overall accuracy obtained for three categories is 65%. The same classifier model was used to test the data set consisting both sentiment and emotion features (SF+EF), that gained the highest accuracy of 93% in ‘negative’ category and the overall accuracy of all three categories is 85%. The highest accuracy obtained against the appropriate sentiment categories are highlighted in bold.

<table>
<thead>
<tr>
<th>SF</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.74</td>
<td>0.89</td>
<td>0.80</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.56</td>
<td>0.53</td>
<td>0.58</td>
</tr>
<tr>
<td>Negative</td>
<td>0.82</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>Mean</td>
<td>0.70</td>
<td>0.73</td>
<td>0.72</td>
</tr>
</tbody>
</table>
The best accuracy obtained by current methods is ranging from 65% to 88%. Souraya Ezzat [14] reported overall accuracy of 74.4 % using SVM Key Graph method. Thus, the proposed method achieved much better accuracy than previous approaches.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, a method has been proposed that accepts a wave format audio clip to extract emotion features and sentiment features. These extracted features used to predict sentiments of a customer speaking to a Call-Centre agent. The extracted three different feature sets are investigated by using well performing classifier. The notable contribution of this work lies in the analysis of the role of human emotions expressed in the form of speech for increasing the intensity of sentiment polarity. Also, it is demonstrated how emotion features can be combined with lexicon based sentiment features in machine learning based sentiment analysis technique. From the experimental results, it is evident that the human emotion signal increases the classification accuracy of sentiment categorization process. The finding of this qualitative study can be used in variety of applications such as automated customer behavior analysis, customer redress systems, customized retail services, and business process quality improvement.

This work can be further extended by correlating the emotion cues with sentiment polarity of corresponding word for exploiting the explicit relationship between sentiment and emotion. Furthermore, this approach can be tested with languages other than English to devise a multi-lingual sentiment analysis system.

REFERENCES


