Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 635 – 643

Procedia
Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining

Jasmine Bhaskar [a,*] ,Sruthi K[a] , Prema Nedungadi [b]

[a]*Department of Compute science , Amrita University, Amrita School of Engineering, Amritapuri, 690525, India*
[b]*Amrita Create, Amrita University, Amrita School of Engineering, Amritapuri, 690525, India*

Abstract

One of the greatest challenges in speech technology is estimating the speaker's emotion. Most of the existing approaches concentrate either on audio or text features. In this work, we propose 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 single feature vector for classification. Our main intention is to increase the accuracy of emotion classification of speech by considering both audio and text features. In this work we use standard methods such as Natural Language Processing, Support Vector Machines, WordNet Affect and SentiWordNet. The dataset for this work have been taken from Semval -2007 and eNTERFACE'05 EMOTION Database.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Emotion classification;text mining;speech mining;hybrid approach

1. 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. Usually emotion

* Corresponding author. Tel.: +91 9447597307;
E-mail address: jasmine@am.amrita.edu

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.112

recognition tasks focus on extracting features from audio. There are different types of temporal and spectral features that can be extracted from human speech. The features like statistics relating to the amplitude and pitch, formants of speech, Mel Frequency Cepstral Coefficients (MFCCs) etc are fed as inputs to the classification algorithms [23]. Speaker's emotion can also be detected using text mining technique on audio material after translating it into text.

Existing human-machine interaction systems can identify "what is said" and "who said it" using speaker identification and speech recognition techniques. These machines can evaluate "how it is said" to respond more correctly and make the interaction more natural, if provided with emotion recognition techniques. Emotion recognition is useful for applications such as Entertainment, e- Learning, and diagnostic tool for therapists, call centre applications etc.

Usually in emotion classification, researchers consider the acoustic features alone. Though features like pitch, energy and speaking rate change with emotional state, strong emotions such as anger and surprise have high pitch and energy. In that case, it is very difficult to distinguish the emotions such as anger and surprise using acoustic features alone. But, if we classify speech solely on its textual component, we will not obtain a clear picture of the emotional content.

In this hybrid approach, we intend to analyze both speech and the corresponding text component in order to detect the speaker's emotion. This method aims to enhance the efficiency of emotion classification by consolidating the features of both audio and text into a single feature vector which is then given to the classifier. The emotional states considered in this experiment are: Happy, Sad, Fear, Disgust, Surprise and Anger. Accordingly the classifier assigns each speaker to one of the above emotional states. Before applying the hybrid approach, both text and speech are handled separately and classified .This allows the comparison of these methods with our proposed hybrid approach.

In the past, some work is done on music mood classification based on lyrics and audio features[1, 2, 3]. The uniqueness of the proposed method is in the choice of the features considered and in the generation of a single feature vector for classification. We propose to use lexical resources like SentiWordNet and WordNet-Affect in order to generate the feature vector for text classification and multi-class SVM for emotional classification.

2. Related Works

This section discusses three aspects of emotion classification: emotion classification from audio, emotion classification from text and emotion classification using both text mining and speech mining

2.1. Emotion classification from the audio

Current research has highlighted new approaches to emotion classification from speech based on audio features. The work by Shen. et .al [21] recognize five different emotion states like disgust, boredom, sadness, neutral and happy using the features: pitch, energy, LPCC, MFCC and LPCMCC. They have explained and compared different combination of features. Their experiment was based on Berlin emotional database. They got different accuracy in different combination of features. In their work [16]Casale.et .al described the working in DSR environment. Features considered for this experiment were extracted by using ETSI ES 202 211 V.1.1.1 S standard front end. In their experiment, they have used two different speech corpora: EMO-DB in German and SUSAS in American English. Their result showed that Support Vector Machine (SVM) trained with Sequential Minimal Optimization (SMO) algorithm leads to better performance.

2.2. Emotion classification from the Text

Mishne[5] worked on classifying blog text according to the mood reported by its author during the writing. Mishne considered different textual features like frequency counts of words, emotional polarity of posts, length of posts, PMI, emphasized words and special symbols like emoticons and punctuation marks. PMI - Point wise Mutual Information provides a numerical weight for keywords based on its relation to a particular mood. SVM (Support Vector Machine) classifier was used in his work for classification. Text mining over transcribed audio recordings was performed in [7], in order to find the speakers emotion. The dataset (audio conversation of the customers) for this experiment was collected from a call centre. The researchers used different feature selection methods in this work. The unsupervised and supervised method further clarified text classification. The evidence demonstrated the effectiveness of the supervised method for text classification. In [8], authors described three different approaches: the statistical approach, the semantic approach and a hybrid statistical-semantic approach. Term-Frequency-Inverse Document Frequency (TF-IDF) values of the terms in the text were used as the features in the statistical approaches. The hybrid statistical–semantic system attained a greater degree of accuracy than either the statistical or the semantic approach.

2.3. Hybrid Approach

The hybrid approach which combines text mining and speech mining is widely used in the field of music genre classification [1,2,3]. In the case of music mood classification, lyrics and audio features were used to improve the accuracy of the classification. In[22] Zhong et al used CHI approach and improved difference based CHI approach to select the features from the lyrics. They have downloaded the mandarin songs from Google music for their testing. Improved CHI approach got better performance than conventional CHI in lyric sentiment task. They have done their classification by combining the audio and text feature by using the serial fusion method. They got better accuracy in fusion method as compared to audio and text features separately. In[4] authors proposed a new emotional classification method for human speech. They have considered a large combination of speech characteristics and text features corresponding to five emotions. For this work, they have developed a dataset of movie quotes. According to their result, hybrid approach attained better accuracy than speech and text mining considered alone.

3. Problem Description

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, we will not obtain a clear picture of the emotional content. In the proposed hybrid approach we consider both text and audio features. Fig. 3 shows the framework for hybrid approach.

4. Methodology

Our approach aims to analyze both speech and the corresponding text component, in order to recognize the emotions in a certain speech. The system automatically classifies speaker's utterances into six emotional states such as happy, surprise, sad, fear, disgust and anger. First, we process both text and audio separately and then the features of audio and text are combined together.

The proposed work consists of three major modules.

- Emotion Classification from speech
- Emotion classification from text
- Hybrid Approach

4.1. Emotion Classification from Speech

In our proposed approach, we extract the common features such as pitch, energy, formants, intensity and ZCR (Zero Crossing Rate). In this work, we extract the formants, zero crossing rate and sound intensity by using Matlab and fundamental frequency (F0), energy using Praat. Fig. 1 shows the framework for emotion classification from speech.

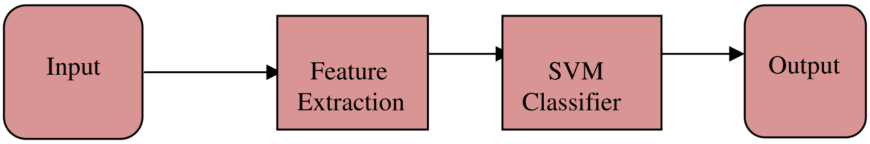

Fig. 1. Emotion Classification from Speech

4.2. Emotion Classification from Text

Emotion classification from text consists of 3 steps.

- Pre-processing
- Emotion detection in the text
- Emotion Classification

Fig 2 shows framework for emotion classification from the text

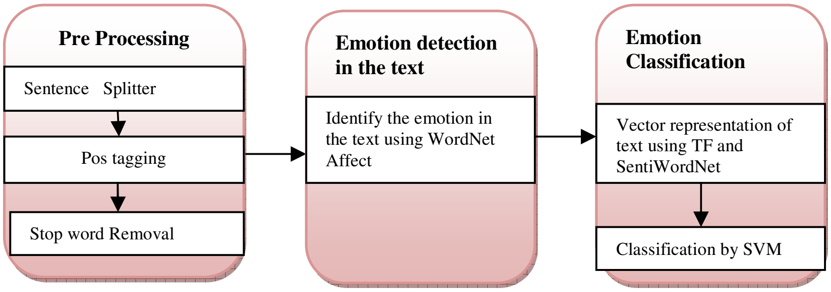
Pre Processing
Sentence Splitter
Pos tagging
Stop word Removal
Emotion detection in the text
Identify the emotion in the text using WordNet Affect
Emotion Classification
Vector representation of text using TF and SentiWordNet
Classification by SVM

Fig. 2 . Emotion Classification from Text

4.2.1 Pre-processing

It follows a similar process to traditional text mining and consists of: sentence splitter, part of speech tagging and stop word removal. First, the document is broken down to sentences. Each word in the sentence is then identified by its corresponding parts of speech. Removal of stop words is the last step in pre-processing. Stop words are the words that don't carry much meaning such as determiners and prepositions. In this work Natural Language tool kit (NLTK) [12] in python is used for pre-processing.

4.2.2. Emotion detection in the text

To identify the emotion in the text we used the English lexical database WordNet Affect[11]. It is a well-known and popular lexical resource, which identifies the emotions that the words convey. WordNet Affect classifies words into six basic emotions: anger, sadness, disgust, fear, happiness and surprise. Algorithm1 is used to identify the emotion in the text. After POS tagging, synonyms set (synset) for each word is obtained from WordNet[11]. WordNet Affect is used to identify the emotional tag of each word. Array tmp[0-5] indicates the frequency count of each emotion. For example, if a sentence contains a sad emotional word then the index containing the count of the emotion `sad' is incremented by one. This step is done for all words in the sentence and their corresponding index is adjusted accordingly. If tmp[0] indicates the number of occurrence of the emotion 'anger' and if it has the maximum value then we can say that the given sentence conveys the emotion 'anger'. After identifying the emotion of each sentence, the next step is emotion classification.

Algorithm 1 Emotion identification in the text

```
 Input : Text
Output: Emotion // happy, disgust, anger, fear, sad, or surprise
    Doc=Filter (Text)
        for each sentence in Doc do
              syns=[] \\ array to store synset
              for  i=0 to 5 do
                    tmp[i]=0 \\ array to store the no: of occurrences of each emotion
```

```
                end for
                for each word in sentence do
                        syns=synset(word) \\ finds and stores the synset from  the WordNet
                end for
                for synonym s in syns do
                        tmp[emotion(s)]++
                end for
                maxemo=max(tmp)
                maxIndex=tmp.Index(maxemo)
        end for
```

4.2.3 Emotion Classification

This module consists of two steps.

- Vector representation of the document
- Emotion classification using SVM.

Algorithm2 describes the emotion feature extraction from the text. In the first step, six dictionaries corresponding to six emotions are created from WordNet Affect. After the parts of speech tagging, synonym set (synset) for each word is determined using WordNet and these words (synset) are stored in array syn. If a word or any of its synonym are present in the dictionaries, then these words are added to an array Emotionwords.

Once all the emotion words in the document are identified, the document is represented as a vector,

Di =$[E_1,E_2,E_3...E_n]$

Where E_i is the emotion word $_i$ with respect to the document. If a term E_1 occurs in a sentence, then it is assigned a non zero value W_i (which denote the weight of that emotion word with respect to the document) otherwise it is zero. Document matrix has each row$_i$ correspond to the feature vector of a particular sentence, and each column of this matrix refers to a unique term (emotion word) in the document.

Weight of the positive emotion word is calculated as W_i=TF_i* Posw$_i$

Weight of the negative emotion word is calculated as W_i=TF_i* Negw$_i$

Here Posw$_i$, Negw$_i$ denote the corresponding scores of the emotion words and is retrieved from the SentiWordNet[6] TF_i is term frequency of the word$_i$ in the dataset.

Algorithm 2 Feature Extraction from Text

```
Input: *Text*
Output: *Feature vector X:*
Dici: Emotion Dictionaries corresponding to six emotions
Emotionwords=[]     \\  array to store emotion words
Doc=Filter(Text)
        for each sentence in Doc do
                syns=[] \\ array to store synset
                for each word in sentence do
                        syns=synset(word) \\ find and stores the synset from  the WordNet
                end for
                for synonym s in syns do
                        if s in Dic$_i$ then
                                Add word to array Emotionwords
                                break;
                        end if
                end for
        end for
        for each word in Emotionwords do
                Find $W_i = TF_i * POSW_i / NEGW_i$
                Store(Emotoword, Wi) as key value pairs
        end for
        for each $sentence_j$ in the Doc do
                for each $word_i$ in the sentence do
                        if word in (Emotoword, Wi) then
                                $X[j][i]=W_i$
                        else
                                X[j][i]=0
                        end if
                end for
        end for
Return feature vector X
```

4.3. Hybrid Approach

The hybrid approach combines both speech mining and text mining to increase the efficiency of emotion recognition of speech. In this approach, features of both text and audio are combined together to form a single feature vector. This single feature vector is then fed as input to the classifier. In this work we have used multi-class SVM for emotion classification. Fig 3 shows the frame work for hybrid approach.

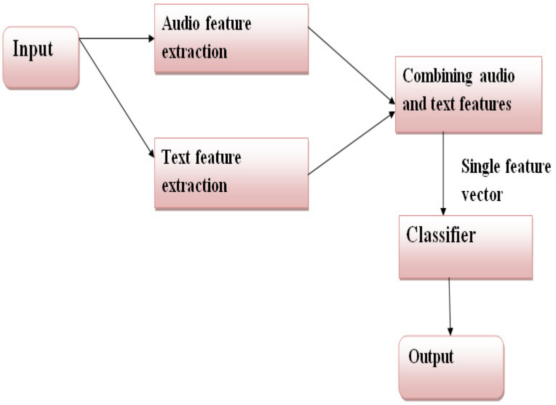

Fig. 3. Hybrid Approach

5. Experiment Results

Three experiments were performed to test the efficiency of our proposed approach: emotion classification using audio features alone, emotion classification using text features alone and classification of combined audio and text features. Two dataset were used for these experiments. The first dataset consisted of news headlines which were taken from SemEval-2007 and Google. As benchmark data was unavailable, wav form of audio of corresponding dataset was developed. We tested the accuracy of speech emotion classification in the first experiment. For this 360 vectors each with 11 features were used to train the SVM classifier. In the second experiment, the accuracy of text emotional classification 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. A third experiment was conducted to calculate the accuracy of the proposed hybrid approach. For this, we combined both speech features and the corresponding text features extracted. 360 vectors each with 96 attributes were used to train the SVM classifier. The accuracy of the above said experiments is given in Table 1. The hybrid approach achieved the highest accuracy and shows better improvement compared to others. The text mining did well compared to speech as most of the emotion words in the dataset are considered in generating the feature vector.

Table 1. Accuracy audio, text and hybrid- Semval-2007

Experiment	Accuracy
Speech emotion Classification	57.1%
Text emotion Classification	76%
Hybrid approach	90%

Table 2 shows the confusion matrix for hybrid approach. Confusion matrix shows the correctly classified emotion count and misclassified emotion count. Each emotion has less misclassification in hybrid approach. Anger, fear, disgust, surprise, happy and sad emotion have seven, five, three, three, two, six misclassification respectively.

Table 2.Confusion Matrix for Hybrid Approach- Semval-2007

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

Second data set used is an audio-visual emotion database named eNTERFACE'05 EMOTION [24], which can be used as a reference database for testing and evaluating video, audio or audio-visual emotion recognition algorithms. We have taken 1530 samples for training and 990 samples for testing. The accuracy of different experiments for this dataset is given in Table 3.

Table 3. Accuracy audio, text and hybrid- eNTERFACE'05

Experiment	Accuracy
Speech emotion Classification	51.1%
Text emotion Classification	62%
Hybrid approach	81%

Table 4 shows the confusion matrix for hybrid approach. Each emotion has less misclassification in hybrid approach. Fear, disgust, surprise and sad emotion have thirty, thirty, thirty six and thirty misclassification respectively. Happy and anger are correctly classified.

Table 4. Confusion Matrix for Hybrid Approach- eNTERFACE'05

Class	Happy	Sad	Fear	Disgust	Surprise	Anger
Happy	165	0	0	0	0	0
Sad	4	135	15	10	0	1
Fear	1	4	135	15	0	0
Disgust	1	5	4	135	10	10
Surprise	3	3	15	10	129	5
Anger	0	0	0	0	0	165

For the second data set also, the hybrid approach achieved the highest accuracy among the three and shows better improvement compared to others. As most of the emotion words in the dataset are considered in generating the feature vector, the text mining did well compared to speech. In the case of audio classification, anger is usually misclassified as surprise and also disgust is misclassified as sad. As the emotional tags of the texts depend on speech and context [4], in such cases text mining also cannot be applied.

6. Conclusion

In this paper we have proposed a new classification method to detect the emotions in utterances of human speech. This method exploits both audio and textual features corresponding to it. In our work we have considered a new combination of audio and text features. Lexicons like WordNet, SentiWordNet and WordNet Affect are used to extract the emotion from the text. Multiclass SVM is used for emotion classification. Our experiment results show that the proposed approach entitles a better accuracy compared to text mining or speech mining. Our findings lead to more realistic human- machine interaction, as it helps to improve the efficiency of emotion classification of human speech. The results show that accuracy and precision of the hybrid approach is significantly higher, compared to the audio or text classification alone.

In this work we have considered only 11 features related to speech mining. In order to enhance the overall performance, more acoustic features can be considered. Further, this work can be linked to many applications like call centers, music recommendation systems and e-learning where speech may replace traditional input devices, in order to make the human-machine interaction in these applications more natural and realistic.

Acknowledgement

This work derives inspiration from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. We thank Dr. Ramachandra Kaimal, Chairman, Department of Computer Science, Amrita University and Dr G J Nair, Department of Electronics and Communication for their support and valuable feedback.

References

1. R Neumayer , A Rauber. Integration of Text and Audio Features for Genre Classification in Music Information Retrieval. In: Advances in Information Retrieval; 2007. p. 724-727.
2. Y Yang, Y Lin, H Cheng, I Liao, Y Ho, H Chen. Toward Multi-Modal Music Emotion Classification. Advances in Multimedia Information Processing- PCM; 2008. p. 70-79.
3. X Hu, J Downie, A Ehmann. Lyric Text Mining in Music Mood Classification. 10th International Symposium on Music Information Retrieval, Kobe; 2009 . p. 411-416.
4. Ali Houjeij, Layla Hamieh, Nader Mehdi, Hazem Hajj. A Novel Approach for Emotion Classification based on Fusion of Text and Speech, 19th international Conference, ICT ; 2012. p. 1-6.
5. G Mishne. Experiments with Mood Classification in Blog Posts. In : Proceedings of ACM SIGIR, 2005 Workshop on Stylistic Analysis of Text for Information Access; 2005.

6. Esuli, F Sebastiani. SentiWordNet A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of 5th International Conference Language Resources and Evaluation, European Language Resources Association, ELRA; 2006. p. 417-422
7. Souraya Ezzat, Neamat El Gayar, Moustafa M Ghanem. Sentiment Analysis of Call Centre Audio Conversations using Text Classification, *International Journal of Computer Information Systems and Industrial Management Applications*, 2012;**4**:619-27.
8. R del Hoyo, I Hupont, F Lacueva, D Abadia. Hybrid Text Affect Sensing System for Emotional Language Analysis. In: Proceedings of the ACM International Workshop on Affective-Aware Virtual Agents and Social Robots: 2009.p. 1-4.
9. D Ververidis, C Kotropoulos. Emotional Speech Recognition Resources, Features, and Methods. *Speech Communication*, 2006;**48**: 1162-81.
10. Iliev, M Scordilis, J Papa, A Falcao. Spoken Emotion Recognition through Optimum-Path Forest Classification Using Glottal Features. *Computer Speech and Language* ; 2010;**24**:445-60.
11. Strapparava, A Valitutti. WordNetAffect an Affective Extension of WordNet. In: Proceedings of LREC, Citeseer;2004.p. 1083-1086.
12. Bird, Steven, Ewan Klein, Edward Loper . *Natural Language Processing with Python*. O'Reilly Media Inc; 2009.
13. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up Sentiment Classification using Machine Learning Techniques In: Proceedings of EMNLP;2002.p. 79-86.
14. T Vogt, E Andre, 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.p. 75-91.
15. R del Hoyo, I Hupont, F Lacueva, D Abadia. Hybrid Text Affect Sensing System for Emotional Language Analysis. In Proceedings of the ACM International Workshop on Affective-Aware Virtual Agents and Social Robots*;* 2009. p. 1-4.
16. S Casale, A Russo, G Scebba, Serrano. Speech Emotion Classification using Machine Learning Algorithms. IEEE International Conference on Semantic Computing;2008.p. 158-165.
17. X Hu , 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.p. 159-168.
18. S Baccianella, A Esuli, F Sebastiani. SentiWordNet 3.0 An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.In: Proceedings of International Conference on Language Resource and Evaluation; 2010.p. 2200-2204.
19. Chihli Hung, Hao-Kai Lin, Chung Yuan.Using Objective Words in SentiWordNet to Improve Word of Mouth Sentiment Classifications. *IEEE Intelligent Systems*, 2013;**28**:47-54.
20. Monalisa Ghosh, Animesh Kar, Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using SentiWordNet 3.0, *International Journal of Engineering Research and Technology 2013*; *2- 9*
21. Peipei Shen, Zhou Chajun, Xiong chen. Automatic Speech Emotion Recognition Using Support Vector Machine. International Conference on Electronic, Mechanical Engineering and Information Technology;2011
22. Jiang Zhong , Yifeng Cheng , Siyuan Yang , Luosheng Wen. Music Sentiment Classification Integrating Audio with Lyrics, *Journal of Information and Computational Science*; 2012;**9**:35-44.
23. Iliou T, Anagnostopoulos. SVM-MLP-PNN Classifiers on Speech Emotion Recognition Field A Comparative Study. Fifth International Conference on Digital Telecommunications ICDT; 2010.p. 1-6.
24. http:// www.enterface.net/enterface05/