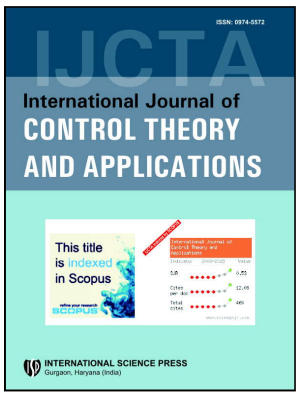

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 18 • 2017

Emotion based Classification of Human Voice using an Optimized Machine Learning Approach

Rabi Narayan Behera[1], Pulak Baral[2], Subindu Saha[2] and Sujata Dash[1]

[1] *Department of Computer Science & Application North Orissa University, Baripada, India, Emails: rabi.behera@iemcal.com,* sujata238dash@gmail.com
[2] *Department of Information Technology, Institute of Engineering & Management, Kolkata, India, Emails: pulakbaral@outlook.com, subindu.saha@gmail.com*

Abstract: Human emotion perception is one of the vast and popular areas of research and it is gaining more and more popularity among researchers, as well as from the many other developing community. There are so many work has been done in area of GA (Genetic Algorithm), K-NN (k-Nearest Neighbors algorithm) and SVMs (Support Vector Machines) in emotion detection in movie review, Twitter blog and over the phone in spoken dialogue systems in different languages especially in English. However, work on perception and production suffers from the difficulty of obtaining reliable human judgments of emotional categories in speech, whether during insight experiments or during corpus labelling. Without reliable exemplars of particular emotional categories, it is difficult to identify those features which contribute to the effect of different emotional categories. In this paper, we divided our 1000 token corpus into training (90%) and test (10%) sets. The test set was randomly selected, but evenly distributed among speakers and tokens. We adopted an ensemble of naïve bayes classifier for binary classification scheme based on the observed ranking distributions to automatically induce prediction models which takes as input training data specifying the class and feature values for each training example and outputs a classification model for use in classifying future examples. R programming is used for data analysis and the proposed ensemble achieves accuracy rivalling the best previously published results.

Keywords: Automatic Speech Recognition, Machine Learning, Emotion Classification, Sound Feature Extraction.

1. INTRODUCTION

In the present scenario, social media and video sharing sites like Facebook and Youtube has become the unlimited source of information of presenting one's opinion. Social media platforms such as Facebook and YouTube are one such big platform where anybody gets the vast information so people visit regularly. Through these videos, people fetch information about sports, movies, social issues, etc. practically speaking Automatic Speech Recognition (ASR)[4] of spontaneous speech is challenging and the resulting transcripts are not very accurate. The difficulty stems from a variety of factors including unknown background environments, noisy audio due to

non-ideal recording conditions, accents, variable source and channel characteristics, various ranges of topics. In our research paper sentiment extraction uses two main systems, ASR (Automatic Speech Recognition) system and text-based sentiment extraction system. For text based sentiment extraction [11], we propose a new method that uses POS (part-of-speech)[15] tagging to extract text features and ensemble based Naïve Bayes uses the text features for prediction of sentiment polarity. Identification of individual text feature's contributions to estimate the sentiment is one of our important features. This provides us with the capability of identifying key words/ phrases within the video that carry important information and can enhance the ability of users to search for appropriate information.

1.1. Automatic Speech Recognition(ASR)

It is an independent automated computerized transcription spoken language of human into comprehensible text format in real time [16]. In a nutshell, it allows a system to identify the words which are spoken by a person and it convert into written text. For a last few decades ASR treated as an important speech research area to understand frequently spoken by a machine Although ASR technology is not accurate to understand all spoken voices, in any acoustic environment, or by any person, it is used on a frequently in a number of applications and services. ASR main goal is to recognize the real time audio with high accuracy, all words that are comprehensibly spoken by any person, independent of noise, vocabulary size, speaker characteristics or accent. For large vocabularies and greater accuracy than 90% system has to be trained for individual speaker's voice. ASR system needs to be trained by individual speaker's voice for a short period of time then it may successfully capture continuous speech and make large vocabulary at normal pace and high accuracy. In optimal condition speech reorganization software can achieve 98% to 99% accuracy it claimed by most commercial companies. Optimal condition usually means that training sample is matched by user speech characteristics, can achieve proper speaker adaptation, and work in a quiet space. Some user's speech is influenced heavily by their mother tongue (accent) so the recognition rates much lower than expected. The fundamental reason behind the research is to improve the accessibility for the deaf and hard of hearing, Searchable text capability and Cost diminution through automation.

1.2. Part of the Speech

PoS tagging is an important problem in the area of Natural Language Processing and plays a key role in most of the machine translation systems which deals with the problem of identification of the pos of each word in the dataset. The PoS tagging process detected the detailed attributes such as nouns, adjective, verbs, number (pronouns) etc... The result of this process is often fed into machine translation systems that use algorithms which make translation decisions based on the tags of the words (part of speech plus attributes). A part of speech tagger can be broken down into two different pipelined components, namely lexical analysis and disambiguation. Lexical based analysis is the initial stage of the Part of speech tagging in which each words are looked up in lexicon dictionaries and assigned with all possible interpretations independent of the context domain, but it suffers from another ambiguous problem with respect to their part of speech for example some English word can be treated as either a noun or a verb in different context or with respect to their attributes like same word (for example read) can be used as either present or past. This problem needs to be solved by the second component in the pipeline i.e. disambiguation.

In this study, we evaluate the proposed sentiment estimation on both publically available text databases and YouTube videos. On the text datasets, our proposed model obtains ~85% accuracy on sentiment polarity detection which is very competitive. The proposed system obtains sentiment polarity of ~85% accuracy On the YouTube videos, which is very encouraging. The paper is prepared as follows: Section 2 covers related work on opinion mining and sentiment recognition from different modalities; Section 3 describes the datasets used and overview of the experiment; next, Sections 4 explains the methods which is for fusing different modalities; finally, Section 5 contain conclusion of the paper and outlines future work.

2. RELATED WORK

Sentiment and emotion analysis both represents a people state of mind; there are only two well known state-of-the-art methods [1, 2] in multimodal sentiment analysis. This section describe the research work done till so far in opinion mining using textual and visual modality. Feature fusion and extraction are crucial for the development of a multimodal sentiment analysis which can be classified into two broad categories: first one feature extraction from each individual modality and second fusion of features technique has to been developed from different textual and visual modality

2.1. Video: sentiment analysis from facial expressions

Emotive analytics is a fascinating mingles of psychology and technology. Many facial expression detection tools classify human emotion into seven main categories like Joy, Sadness, Anger, Contempt, Fear, Surprise and Disgust. The Active appearance model [3,4] and Optical flow based technique[5] are common approaches that use FACS to understand facial expressions expressed by humans in real time. Exploiting action units as features like Bayesian networks, Hidden Markov models (HMM), k-nearest neighbour and artificial neural networks (ANN) has helped many researchers to infer emotions from facial expressions by using various manually crafted corpora, which make it nearly impossible for a comparative performance evaluation.

2.2. Audio: sentiment recognition from speech

Recent research on speech-based emotion analysis [4,7–10] have focused on identifying several audio features such as pitch(frequency), intensity of utterance [11], bandwidth, and audio duration. Navas et al. [12] has claimed that the speaker-dependent approach performs better than the speaker-independent approach, where about 98% accuracy was achieved by using the Gaussian mixture model (GMM) as a classifier, with speech features like prosodic voice quality as well as Mel frequency cepstral coefficients (MFCC). The speaker-dependent approach is not reasonable in many applications that deal with a very large number of possible users (speakers).

The task of multimodal sentiment analysis is addressed by Rosas, Mihalcea and Morency [14] presents a method that integrates linguistic, visual and audio features for the purpose of sentiment identification in online Spanish videos collected from the different social media website and annotated for sentiment polarity.

They experimentally proved that the combined use of audio, visual and textual features improves over the use of one modality at a time. They used Support Vector Machines (SVM) with a linear kernel, which are binary classifiers that seek to find the hyper plane that best separates a set of positive examples from a set of negative examples, with maximum margin. They use the uses a ten-fold cross validation on the entire dataset by WEKA machine learning toolkit for each experiment.

Form the above related work, for speaker-independent applications, the best classification accuracy achieved [13, 14], obtained on the Berlin Database of Emotional Speech and Spanish videos , which motivate us to apply a new approach, by combining ensemble based classifier with fusion of features coming from different modalities provides very encouraging outcome.

3. DATASET EMPLOYED

From YouTube twenty videos were collected related to different topics like sports, politics, movie reviews, etc. The videos were found using the following keywords: opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, I hate, I like. The final video set had 10 female and 10 male speakers randomly selected from YouTube, with their age ranging approximately from 14 to 60 years. Although they belonged to different ethnic backgrounds (e.g., Caucasian, African-American, Hispanic, Asian), all speakers expressed themselves in English. The videos were converted to mp4 format with a standard size of 360 480. The

length of the videos varied from 2 to 5 min. All videos were pre-processed to avoid the issues of introductory titles and multiple topics. Each provided video segmented, transcribed and labeled by a sentiment value. Because of this annotation scheme of the dataset, textual data was available for our experiment. We used multimodal sentiment analysis system and to evaluate the system's performance.

We train a Naive Bayes Classifier using a trained corpus taken from YouTube video. This could be improved using a better training by employing ensemble based majority voting techniques.

3. OVERVIEW OF THE EXPERIMENT

In order to analyze the voice sentiments, Final feature vector contain the information of audio which is formed by merging audio, visual and textual feature vector. Later a supervised classifier was employed on the fused feature vector to identify the overall polarity of each segment of the video clip. We also carried out an experiment on decision-level fusion, by taking the sentiment classification result from three individual modalities as inputs and produced the final sentiment label.

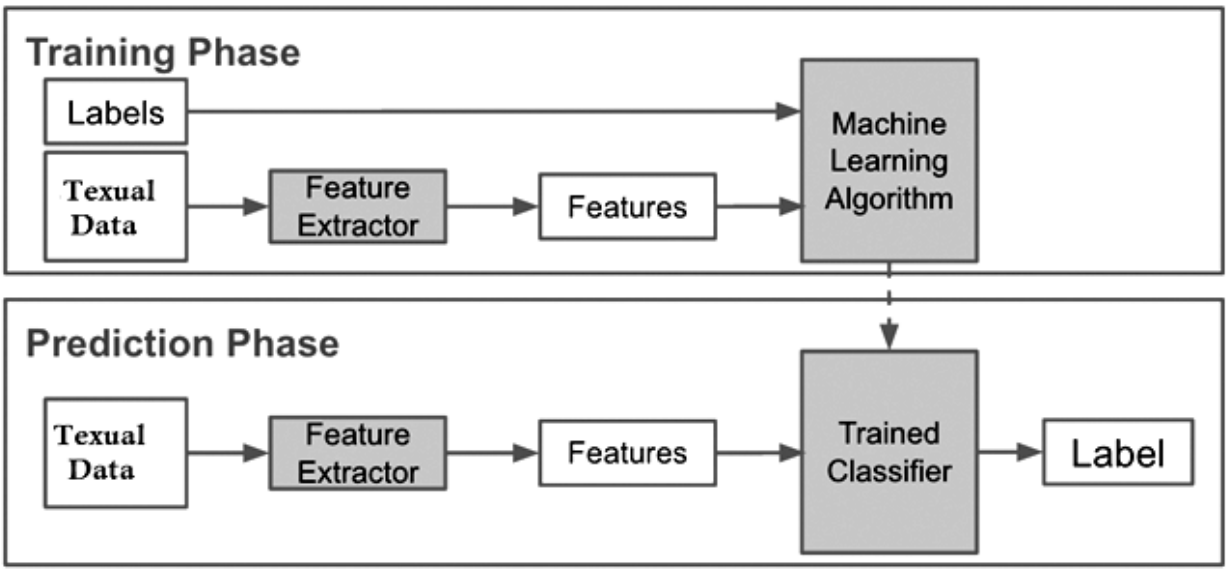

Figure 1: Machine Learning Phase

On the same training and test sets, we ran the classification experiment using multiple Naïve Bayes classifier. This trivial method had the advantage of relative simplicity, yet was shown to produce significantly high accuracy. We concatenated the feature vector of each modality into a single stream. This feature vector was then used for classifying each video segment into sentiment classes. To estimate the accuracy, we used 10-fold cross validation. The result is shown in the table1.

Table 1
Results of feature-level fusion

Combination of modalities	*Precision*	*Recall*	*Accuracy from previous work[14]*
Accuracy of the experiment carried out on Textual Modality	0.821	0.63	54.05%
Accuracy of the experiment carried out on Audio Modality	0.852	0.65	48.64 %
Result obtained using audio and text-based features	0.84	0.83	64.86%

Our result is compared with the result obtained previously [14] on Multimodal Sentiment Analysis of Spanish Online Videos and found satisfactory.

4. CONCLUSION AND FUTURE WORK

In this paper we presented an ensemble based machine learning sentiment analysis framework, which includes sets of relevant features for textual, audio and visual data. In particular, our textual sentiment analysis module has been enriched by sentic-computing-based features, which have offered substantial upgrading in the performance of our model. Our future work will aim to include gaze and smile features, sentiment classification from facial-expression-based, in addition to focusing on the use of audio modality for the multimodal sentiment classification task. Furthermore, we will explore the possibility of developing a culture- and language independent multimodal sentiment classification framework.

REFERENCES

[1] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: harvesting opinions from the web, In: Proceedings of the 13th International Conference on Multimodal Interfaces, ACM, Alicante, Spain, 2011, pp. 169–176.

[2] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: scalable multimodal fusion for continuous interpretation of semantics and sentics, In: IEEE SSCI, Singapore, 2013, pp. 108–117.

[3] A. Lanitis, C.J. Taylor, T.F. Cootes, A unified approach to coding and interpreting face images, In: Fifth International Conference on Computer Vision, 1995. Proceedings, IEEE, Cambridge, Massachusetts, USA, 1995, pp. 368–373.

[4] D. Datcu, L. Rothkrantz, Semantic audio–visual data fusion for automatic emotion recognition, In: Euromedia, Citeseer, 2008.

[5] M. Kenji, Recognition of facial expression from optical flow, IEICE Trans. Inf. Syst. 74 (10) (1991) 3474–3483.

[6] N. Ueki, S. Morishima, H. Yamada, H. Harashima, Expression analysis/synthesis system based on emotion space constructed by multilayered neural network, Syst. Comput. Jpn. 25 (13) (1994) 95–107

[7] I.R. Murray, J.L. Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, J. Acoust. Soc. Am. 93 (2) (1993) 1097–1108.

[8] R. Cowie, E. Douglas-Cowie, Automatic statistical analysis of the signal and prosodic signs of emotion in speech, In: Fourth International Conference on Spoken Language, 1996. ICSLP 96, Proceedings, vol. 3, IEEE, Philadelphia, PA, USA, 1996, pp. 1989–1992.

[9] F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech, In: Fourth International Conference on Spoken Language, 1996, ICSLP 96, Proceedings, vol. 3, IEEE, Philadelphia, PA, USA, 1996, pp. 1970–1973.

[10] T. Johnstone, Emotional speech elicited using computer games, In: Fourth International Conference on Spoken Language, 1996, ICSLP 96, Proceedings, vol. 3, IEEE, Philadelphia, PA, USA, 1996, pp. 1985–1988.

[11] L.S.-H. Chen, Joint processing of audio–visual information for the recognition of emotional expressions in human–computer interaction (Ph.D. thesis), Citeseer, 2000.

[12] E. Navas, I. Hernáez, I. Luengo, An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional tts, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006) 1117–1127.

[13] H. Atassi, A. Esposito, A speaker independent approach to the classification of emotional vocal expressions, In: ICTAI, 2008, pp. 147–150.

[14] Rosas, Veronica, Rada Mihalcea, and Louis-Philippe Morency. "Multimodal sentiment analysis of Spanish online videos." IEEE Intelligent Systems 28.3 (2013): 38-45.

[15] Nicholls, Chris, and Fei Song. "Improving sentiment analysis with part-of-speech weighting." 2009 International Conference on Machine Learning and Cybernetics. 2009.

[16] Stuckless, R." Developments in real-time speech-to-text communication for people with impaired hearing". In M. Ross(Ed.), Communication access for people with hearing loss (pp.197-226). Baltimore, MD: York Press(1994).