Multimodal Sentiment Analysis Using Deep Neural Networks

Harika Abburi[1(✉)], Rajendra Prasath[2], Manish Shrivastava[1], and Suryakanth V. Gangashetty[1]

[1] Langauage Technology Research Center,
International Institute of Information Technology Hyderabad, Hyderabad, India
harika.abburi@research.iiit.ac.in, {m.shrivastava,svg}@iiit.ac.in
[2] NTNU, Trondheim, Norway
drrprasath@gmail.com

Abstract. Due to increase of online product reviews posted daily through various modalities such as video, audio and text, sentimental analysis has gained huge attention. Recent developments in web technologies have also enabled the increase of web content in Hindi. In this paper, an approach to detect the sentiment of an online Hindi product reviews based on its multi-modality natures (audio and text) is presented. For each audio input, Mel Frequency Cepstral Coefficients (MFCC) features are extracted. These features are used to develop a sentiment models using Gaussian Mixture Models (GMM) and Deep Neural Network (DNN) classifiers. From results, it is observed that DNN classifier gives better results compare to GMM. Further textual features are extracted from the transcript of the audio input by using 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 an online product reviews.

Keywords: Multimodal sentiment analysis · MFCC · Doc2Vec · GMM · SVM · Deep neural networks

1 Introduction

Based on the opinion, Sentiment analysis classifies data into positive, negative and neutral categories. As of now most of the work on sentiment analysis is done on textual data. With increase in social media, people started sharing the information in the form of video, audio along with text. So all kinds of data are required for better sentiment classification. For any kind of approach like audio and text, sentiment can be extracted using sentiment classification techniques like lexicon based approach and machine learning approach [8].

R. Prasath—A part of this was carried out when the author was in Indian Institute of Information Technology (IIIT) Sricity, India.

R. Prasath and A. Gelbukh (Eds.): MIKE 2016, LNAI 10089, pp. 58–65, 2017.
DOI: 10.1007/978-3-319-58130-9_6

Audio sentiment detection system is developed on the lines of Maximum Entropy modeling and Part Of Speech tagging. Transcripts from audio streams are obtained using Automatic Speech Recognition (ASR) [3]. In [4] rather than using ASR, Key Word Spotting (KWS) is used to extract the sentiment. Experiments have shown that the presented approach outperforms the traditional ASR. Authors of [6] have worked on determining if prosodic features can be used to build the sentiment classifier. Speech data which is there in audio file are generally extracted from the vocal track, excitation and prosody. Audio features like pitch, intensity and loudness are extracted using OpenEAR software and Support Vector Machine (SVM) classifier is built to detect the sentiment [12]. The audio features are automatically extracted from the audio track of each video clip using OpenEAR software and Hidden Markov Models (HMM) classifier is built to detect the sentiment [9]. In this paper Mel Frequency Cepstral Coefficients (MFCC) features are extracted and tested using DNN and GMM classifier.

Movie review mining using machine learning and semantic orientation is implemented [1]. In semantic orientation approach, bad and good associations account for negative and positive and based on the document features we can classify whether an input review belongs to a negative or positive class. Machine learning techniques are used to investigate the effectiveness of classification of documents by overall sentiment [10]. A variety of features like unigrams, bigrams and combination of both were employed, but the best results came from unigrams run through an SVM. Sentiment analyzer is developed to find out all the references on the subject and the sentiment polarity of each reference from online data documents [17]. To develop the sentiment analyzer, sentiment lexicon and the sentiment pattern database is used for extraction and association purposes. They classify expressions about specific items and use manually developed patterns to classify polarity. These patterns are high-quality, yielding a quite high precision, but very low recall. Sentiment is extracted using the opinion words like a combination of the adjectives along with the verbs and adverbs in the tweets [5]. They preprocess the tweets and add weightage according to the number of exclamation marks and the adjectives, verbs and adverbs are tagged in each tweet. Adjectives and negative words are taken into account to calculate the polarity of the whole phrase. The corpus-based method was used to find the semantic orientation of adjectives and the dictionary-based method to find the semantic orientation of verbs and adverbs. Two different Naive Bayes classifiers which make use of polarity lexicon to classif as positive and negative are used to detect the polarity of English tweets [2]. These classifiers are treated as the baseline. Features like lemmas, multiword, polarity lexicon and valence shifters are used. The training data set of tweets is obtained from SemEval 2014 and additional annotated tweets from external sources. Experiments show that performance is best when binary strategy is used with multiword and valence shifters. Approach to analyze the sentiment of short Chinese texts is presented in [15]. By using word2vec tool, sentiment dictionaries from NTU and HowNet are extended. Then the feature weight of the words are enhanced including the

words that appear in the sentiment dictionary and the words next to the sentiment words. The model is implemented using SVM classifier.

The joint use of multiple modalities such as video, audio and text is explored for developing a sentimental model. Both feature level and decision level fusion methods are used to merge effective information extracted from multiple modalities. An improvement over classification by grouping over different modalities is reported in [14]. Multimodal sentiment analysis approach is an intelligent opinion mining system for identifying and understanding sentiment present in the reviews. In order to extract the sentiment they used audio and video signals and hence overcome the drawbacks of traditional sentiment analysis system [16]. In [7], the authors collected the English dataset from YouTube and expotv. The feature basis is formed by using text, video and audio features. Based on the textual movie review corpus, different levels of domain-dependence are considered such as in-domain analysis and cross-domain analysis. This shows that cross-corpus training works sufficiently well. Authors of [11] introduce database consisting of Spanish videos. They explored the combination of three modalities such as text, speech and video features on classification. They even explore the same work in English videos. From the results it is observed that the joint use of three modalities bring significant improvement. To determine the sentiment polarity present in the input [13] extracted the features from three modalities. The convolution neural network is used to extract the text features showed significant improvement in detecting the sentiment from a review.

In this work, a method to combine both the text and audio features is explored to detect the sentiment from the online product reviews. As of now, less research is done on the multimodal sentiment analysis of online reviews in Hindi language. Our proposed system is implemented in Hindi database. From the literature, it is observed that sentiment is detected from the input by extracting several features using OPENEAR/OPENSMILE tool and build a system using the SVM classifier. Instead of taking all the features, in this paper MFCC 13 dimension feature vector is extracted from each audio input and build sentiment analysis system using GMM and DNN classifiers. In order to detect the sentiment of a text input, textual features which are computed by Doc2Vec vectors are used to build the SVM classifier.

The rest of the paper is organized as follows: Hindi product reviews database used in this paper is discussed in Sect. 2. Sentiment analysis using audio features is discussed in Sect. 3. Sentiment analysis using text features is discussed in Sect. 4. Multimodal sentiment analysis and experimental results of proposed method for detecting the sentiment of a Hindi data is discussed in Sect. 5. Finally, Sect. 6 concludes the paper with a mention on the future scope of the present work.

2 Hindi Database

The database used in our studies is collected from YouTube, which is a publicly available source. The dataset includes reviews of phones, lotions and shampoos.

The database has some degree of generality as a variety of product reviews are used within the broad domain of product reviews. The two basic sentiments presented in the database are: Positive and Negative. The average length of each input is thirty seconds and average number of words in each input is around 40. A total of 110 product reviews are collected, among them 100 inputs are taken based on inter-annotator agreement. Transcription and sentiment annotations were manually performed for text based sentiment classification. Both the modalities such as audio and text are provided for annotators to figure out the exact opinion of the input. Then based on inter-annotator agreement, 50 positive and 50 negative inputs are selected. Among them 80% are used for training and remaining 20% are used for testing.

3 Sentiment Analysis Using Audio Features

The process of developing the sentiment model and extracting the audio features from the input is described in this section. The features which are extracted are used to build a classifier of positive or negative sentiment. Each input is in the form of .wav format, in 16 bit, 16000 Hz sampling frequency and a mono channel. MFCC features are extracted from each input and sentiment model is developed using Gaussian Mixture Models (GMM) and Deep Neural Network classifiers. Block diagram of sentiment analysis using Deep Neural Network is shown in the Fig. 1.

A deep neural network (DNN) is a neural network with multiple hidden layers of nodes between the input and output layers. These hidden layers do feature identification and processing in a series of stages. The successive layers can learn higher level features. DNN performance depends on training data. The more the training data the more accurate it was. Each DNN is trained for 30 epochs with different number of layers and different number of nodes in each layer. In our work up to four deep layers are explored. All the DNNs are trained with ADAM method which is hyper parameter learning algorithm. MFCC features considered in this study are 13-Dimension, 65-Dimension and 130-Dimension. 10 frames of 13 dimension MFCC frames are concatenated to get 130 dimension and 5 frames of 13 dimension MFCC frames are concatenated to get 65 dimension. Frames are concatenated here because each frame will not carry the sentiment. So, experiments are even done in combination of frames, which results in better performance. Based on the input dimension, the input layer nodes can be 13, 65 and 130 which are linear. The output layer is of softmax layer with 2 nodes because the number of classes in the database are 2. During testing the node which gives maximum score is assigned as the claimed class. For testing only 5 s of data is taken for each input.

MFCC features which are extracted from each input is given as input to the GMM. GMM is tested with different number of test cases and different number of mixtures like 16, 32 and 64. Here also for testing 5 s of data is taken.

From the Table 1 it is observed that DNN with 65-Dimension feature vector has performed better when compared to 130 dimensions because by using 130

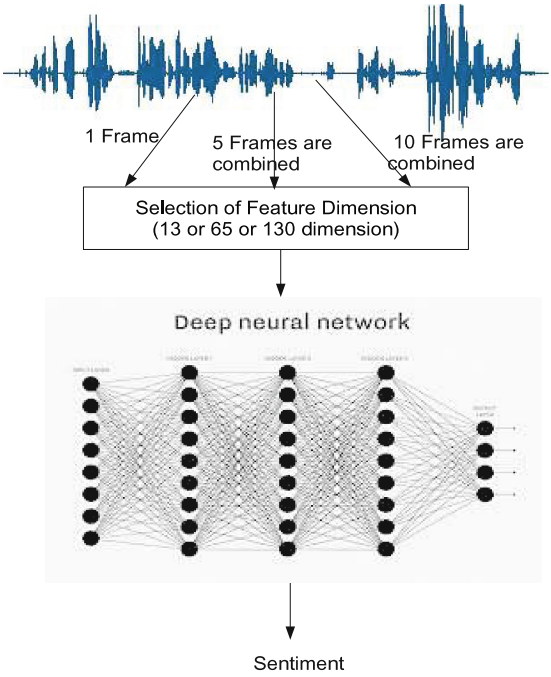

Fig. 1. Block diagram of sentiment analysis using deep neural network

Table 1. Performance of sentiment analysis using deep neural network

DNN	Two layers (%)	Three layers (%)
13-Dimension	41.66	58.3
65-Dimension	58.3	**75.0**
130-Dimension	58.3	66.7

Table 2. Performance of sentiment analysis using classifiers

Classifier	Accuracy (%)
GMM (64)	58.3
DNN (three layer)	**75.0**

dimension feature vector, features are not sufficient to train the DNN. It is also observed that performance is more with three layers. The fourth layer is also explored but we are getting same accuracy as in the third layer. From the Table 2 it is observed that DNN with three hidden layers outperforms the GMM with 64 mixtures by 17%.

4 Sentiment Analysis Using Text Features

The process of developing the sentiment model and extracting the text features from the input is described in this section. These features are used to build a classifier of positive or negative sentiment. In a preprocessing step, each audio input is manually transcribed and sentiment annotations are also assigned manually. For better results, 300 dimension feature vector is generated from each text input.

Many machine learning algorithms require the input to be represented as a fixed-length feature vector. Word2Vec takes data from a corpus, and churns out vectors for each of those words. These vectors are interesting because similar words are placed nearer to each other in the vector space. Doc2Vec model represents not only words, but it is an unsupervised learning of continuous representations for larger text such as sentences, paragraphs and whole documents. In Doc2Vec architecture, the algorithms used are distributed memory and distributed bag of words. Distributed memory will randomly initialize paragraph vector for each document and predict next instance using context words and paragraph vectors. Context window slide across document, but paragraph vector is fixed. On the other hand distributed bag of words will only use paragraph vectors and not word vectors. It will take window of words in a paragraph and randomly sample which one to predict using paragraph vector (ignores word ordering). By combining both the algorithms Doc2Vec generates the vectors.

Doc2Vec will generate a single vector for each manually transcribed input, which represents the meaning of a document. To associate documents with labels this vector will be used as an input to a supervised machine learning algorithm. Sentiment analysis based on text can be viewed as a text classification task which can be handled by SVM. SVM classifier is trained with vectors generated from the doc2vec and by using corresponding sentiment tags as positive or negative. Given a test input, the trained models classify it as either positive or negative. From Table 3 it is observed that the rate of detecting the sentiment of Hindi language product reviews for text features is 63.64 %.

Table 3. Performance of sentiment analysis using text features

Classifier	Accuracy (%)
SVM	63.64

5 Multimodal Sentiment Analysis

Analyzing the audio data has advantage of voice modularity compared to the textual data. In textual data, exact sentiment of the input may not extract properly because it only has the information regarding the words and their dependencies. Instead, audio data contain multiple modalities like acoustic and

Table 4. Performance of multimodal sentiment analysis

Modality	Accuracy (%)
Audio	75
Text	63.4
Audio + text	**78.2**

linguistic streams. Both the modalities are hypothesized based on the highest average probability of the classifiers. From our experiments, it is observed that the simultaneous use of these two modalities help to create a better sentiment analysis model to detect whether the given test input is positive or negative sentiment.

From the Table 4 it is observed that by combining the two modalities such as text and audio rate of detecting the sentiment of a product review is improved.

6 Conclusion

In this paper, we proposed an approach to extract the sentiment of a given input using both audio and text information. MFCC features are extracted from audio and sentiment models are built using DNN and GMM. DNN is tested with different layers with different number of nodes, whereas GMM is tested with different mixture components with different test cases. For text, features generated using Doc2Vec are used to build the model using the SVM classifier. From our experiments, it is observed that DNN classifier with 65-dimension MFCC features has high accuracy of detecting the sentiment of an input compared to other dimensions and even DNN classifier outperformed GMM classifier by 17%. It is also observed that by combining both the modalities such as audio and text, rate of detecting the sentiment is significantly improved.

References

1. Chaovalit, P., Zhou, L.: Movie review mining: a comparison between supervised and unsupervised classification approaches. In: Proceedings of IEEE 38th Hawaii International Conference on System Sciences, Big Island, Hawaii, pp. 1–9 (2005)
2. Gamallo, P., Garcia, M.: Citius: a naive-bayes strategy for sentiment analysis on english tweets. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 171–175, August 2014
3. Kaushik, L., Sangwan, A., Hansen, J.H.L.: Sentiment extraction from natural audio streams. In: Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 8485–8489 (2013)
4. Kaushik, L., Sangwan, A., Hansen, J.H.: Automatic audio sentiment extraction using keyword spotting. In: Proceedings of Interspeech, pp. 2709–2713, September 2015
5. Kumar, A., Sebastian, T.M.: Sentiment analysis on twitter. IJCSI Int. J. Comput. Sci. **9**(4), 372–378 (2012)

6. Mairesse, F., Polifroni, J., Fabbrizio, G.D.: Can prosody inform sentiment analysis? experiments on short spoken reviews. In: Proceedings of IEEE International Confernce on Acoustics, Speech and Signal processing (ICASSP), pp. 5093–5096 (2012)
7. Wollmer, M., Felix, W., Knaup, T., Morency, L.P.: YouTube movie reviews: sentiment analysis in an audio-visual context. IEEE Intll. Syst. **28**(3), 46–53 (2013)
8. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**, 1093–1113 (2014)
9. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI2011), pp. 169–176, November 2011
10. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
11. Perez-Rosas, V., Mihalcea, R., Morency, L.P.: Multimodal sentiment analysis of spanish online videos. IEEE Intll. Syst. **28**(3), 38–45 (2013)
12. Perez-Rosas, V., Mihalcea, R., Morency, L.P.: Utterance level multimodal sentiment analysis. In: Proceedings of ACL, pp. 973–982 (2013)
13. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of EMNLP, pp. 2539–2544 (2015)
14. Poria, S., Cambria, E., Howard, N., Huang, G.B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing **174**, 50–59 (2015)
15. Xing, L., Yuan, L., Qinglin, W., Yu, L.: An approach to sentiment analysis of short chinese texts based on SVMs. In: Proceedings of the 34th Chinese Control Conference, pp. 28–30. IEEE, July 2015
16. Yadav, S.K., Bhushan, M., Gupta, S.: Multimodal sentiment analysis: sentiment analysis using audiovisual format. In: Proceedings of IEEE 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1415–1419 (2015)
17. Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.: Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of IEEE International Conference on Data Mining (ICDM) (2003)