Multimodal Sentiment Analysis of Telugu Songs

by

Harika Abburi, Eashwar Sai Akhil, Suryakanth V Gangashetty, Radhika Mamidi

Hilton, New York City, USA.

Report No: IIIT/TR/2016/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2016

Multimodal Sentiment Analysis of Telugu Songs

Harika Abburi, Eswar Sai Akhil Akkireddy, Suryakanth V Gangashetty, Radhika Mamidi
Langauage Technology Research Center
IIIT Hyderabad India
{harika.abburi, eswarsai.akhil}@research.iiit.ac.in
{svg, radhika.mamidi}@iiit.ac.in

Abstract

In this paper, an approach to detect the sentiment of a song based on its multi-modality natures (text and audio) is presented. The textual lyric features are extracted from the bag of words. By using these features, Doc2Vec will generate a single vector for each song. Support Vector Machine (SVM), Naive Bayes (NB) and a combination of both these classifiers are developed to classify the sentiment using the textual lyric features. Audio features are used as an add-on to the lyrical ones which include prosody features, temporal features, spectral features, tempo and chroma features. Gaussian Mixture Models (GMM), SVM and a combination of both these classifiers are developed to classify the sentiment using audio features. GMM are known for capturing the distribution in the features and SVM are known for discriminating the features. Hence these models are combined to improve the performance of sentiment analysis. Performance is further improved by combining the text and audio feature domains. These text and audio features are extracted at the beginning, ending and for the whole song. From our experimental results, it is observed that the first 30 seconds(s) of a song gives better performance for detecting the sentiment of the song rather than the last 30s or from the whole song.

1 Introduction

Sentiment analysis is defined as a task of finding the opinion about specific entities. In our case it is a task of finding the sentiment of a song. With the growing amount of music and the demand of human to access the music information retrieval, music sentiment analysis is emerging as an important and essential task for various system and applications. To extract the sentiment, thousands of text, audio and video documents will process in few seconds . Sentiment analysis mainly focuses on two approaches, text based and audio based [Tyagi and Chandra, 2015]. For any approach sentiment can be extracted using sentiment classification techniques like machine learning approach, lexicon based approach and hybrid approach [Medhat *et al.*, 2014].

In lyric-based song sentiment classification, sentiment-vector space model is used for song sentiment classification [Xia *et al.*, 2008]. Experiments are done on two approaches: knowledge-based and machine learning. In knowledge-based, HowNet [Dong *et al.*, 2010] is used to detect the sentiment words and to locate the sentiment units with in the song lyric. In machine learning, the SVM algorithm is implemented based on Vector Space Model (VSM) and sentiment-Vector Space Model (s-VSM), respectively. Experiments show that s-VSM gives better results compared to VSM and knowledge-based. A previous work includes sentiment analysis for mining the topics from songs based on their moods [Shanmugapriya and Dr.B.Srinivasan, 2015]. The input lyrics files are measured based on the wordnet graph representation and the sentiments of each song are mined using Hidden Markov Model (HMM). Based on single adjective words available from the audio dataset USPOP, a new dataset is derived from the last.fm tags [Hu *et al.*, 2007]. Using this dataset, K-means clustering method is applied to create a meaningful cluster-based set of high-level mood categories for music mood classification. This set was not adopted by others because mood categories developed by them were seen as a domain oversimplification. The authors in [Hu *et al.*, 2009] presented the usefulness of text features in music mood classification on 18 mood categories derived from user tags and they show that these text features outperform audio features in categories where samples are more sparse. An unsupervised method to classify music by mood is proposed in [Patra *et al.*, 2013]. Fuzzy c-means classifier is used to do the automatic mood classification.

In audio-based song sentiment classification: A method is presented for audio sentiment detection based on KeyWord Spotting (KWS) rather than using Automatic Speech Recognition (ASR) [Kaushik *et al.*, 2015]. Experiments show that the presented method outperform the traditional ASR approach by 12 percent increase in classification accuracy. Another method for detecting the sentiment from natural audio streams is presented [Kaushik *et al.*, 2013]. To obtain the transcripts from the video, ASR is used. Then a sentiment detection system based on Maximum Entropy modeling and Part of Speech tagging is used to measure the sentiment of the transcript. The approach shows that it is possible to automatically detect sentiment in natural spontaneous audio with good accuracy. Instead of using KWS and ASR we can di-

rectly extract the features like prosody, spectral etc to detect the sentiment of a song from audio. For music audio classification, instead of using Mel Frequency Cepstral Coefficients (MFCC) and chroma features separately combination of both gives better performance. Because chroma features are less informative for classes such as artist, but contain information which is independent of the spectral features [Ellis, 2007]. Due to this reason in our work, experiments are done by combining both features along with some other features.

Instead of using only lyrics or only audio, research is also done on combinations of both the domains. In [Hu and Downie, 2010] work is done on the mood classification in music digital libraries by combining lyrics and audio features and discovered that complementing audio with lyrics could reduce the number of training samples required to achieve the same or better performance than single source-based systems. Music sentiment classification using both lyrics and audio is presented [Zhong *et al.*, 2012]. For lyric sentiment classification task, CHI approach and an improved difference-based CHI approach were developed to extract discriminative affective words from lyrics text. Difference-based CHI approach gives good results compare to CHI approach. For audio sentiment classification task, features like chroma, spectral etc. are used to build SVM classifier. Experiments show that the fusion approach using data sources help to improve music sentiment classification. In [Jamdar *et al.*, 2015], [Wang *et al.*, 2009] music is retrieved based on both lyrics and melody information. For lyrics, keyword spotting is used and for melody MFCC and Pitch features are extracted. Experiments show that by combining both modalities the performance is increased.

In this work, a method to combine both lyrics and audio features is explored for sentiment analysis of songs. As of now, less research is done on multimodal classification of songs in Indian languages. Our proposed system is implemented on Telugu database. For lyrics, Doc2Vec is used to extract the fixed dimension feature vectors of each song. SVM and Naive Bayes classifiers are built to detect the sentiment of a song due to their excellence in text classification task. For audio, several features are extracted like prosody, temporal, spectral, chroma, harmonics and tempo. Classifiers that are built to detect the sentiment of a song are SVM, GMM and combination of both. It is observed that in the literature a lot of work is done on whole song to know the sentiment, but the whole song will not give good accuracy because the whole song may or may not carry the same attribute like happy (positive) and sad (negative). The beginning and the ending parts of the song includes the main attribute of that song. Hence, experiments are done on different parts of the song to extract the sentiment.

The rest of the paper is organized as follows:Database and classifiers used in this work is discussed in section 2 and sentiment analysis using lyric features is discussed in section 3. Sentiment analysis using audio features is discussed in section 4. Multimodal sentiment analysis and experimental results in proposed method for detecting the sentiment of a song is discussed in section 5. Finally, section 6 concludes the paper with a mention on the future scope of the present work.

2 Database and Classifiers used in this study

The database used in this paper is collected from the Youtube which is a publicly available source. A total of 300 Telugu movie songs and lyrics corresponding to each song are taken. The two basic sentiments presented in the database are: Happy and Sad. Joyful, thrilled, powerful, etc are taken as happy sentiment and ignored, depressed, worry, etc are taken as sad sentiment. As our native language is Telugu, work is implemented on Telugu songs which don't have any special features compared to other language songs. Telugu songs are one of the popular categories of Indian songs and are present in Tollywood movies. Most of the people belonging to the south part of India will listen to these songs.The songs include variety of instruments along with the vocals. Here the main challenging issue is the diversity of instruments and vocals. The average length of each song is three minutes thirty seconds and average number of words in lyrics for each song is around 300. The database is annotated for the sentiment happy and sad by three people. Annotators are provided with the two modalities such as text and audio to correctly figure out the sentiment of a song. Then based on inter-annotator agreement, 50 happy songs and 50 sad songs are selected because some songs seems to be happy or sad for one annotator and neutral to another annotator. So, only 100 songs are selected out of 300. Inter-annotator agreement is a measure of how well two or more annotators can make the same annotation decision for a certain category. Among them 40% of songs are used for training and 60% of songs are used for testing.

2.1 Naive Bayes

Naive Bayes classifier is a probabilistic classifier of words based on the Bayes theorem with an independence assumption that words are conditionally independent of each other. This assumption does not affect the accuracy in text classification but makes really fast classification algorithm. Despite the assumptions that this technique uses, Naive Bayes performs well in many complex real-world problems. Multinomial Naive Bayes is used in our system where the multiple occurrences of the words matter a lot in the classification problem.

The main theoretical drawback of Naive Bayes method is that it assumes conditional independence among the linguistic features. If the main features are the tokens extracted from texts, it is evident that they cannot be considered as independent, since words co-occurring in a text are somehow linked by different types of syntactic and semantic dependencies. Despite its simplicity and conditional independence assumption, Naive Bayes still tends to perform surprisingly well [Rish, 2001]. On the other hand, more sophisticated algorithms might yield better results; such as SVM.

2.2 Support Vector Machines

Support vector machine classifier is intended to solve two class classification problems. The basic principle implemented in a support vector machine is that the input vectors which are not linearly separable are transformed to a higher dimensional space and an optimum liner hyperplane is designed to classify both the classes. An SVM [Campbel *et al.*,

2006] is a two-class classifier constructed from sums of a kernel functions.

2.3 Gaussian Mixture Models

GMMs are well known to capture the distribution of data in the feature space. A Gaussian mixture density is a sum of M weighted component densities [Reynolds and Rose, 1995] given by the equation:

$$p(x_k|\lambda) = \sum_{r=1}^{M} w_r K_r(x_k) \quad (1)$$

where x_k is an N dimensional input vector,
$K_r(x_k), r = 1...M$ are the component densities and
$w_r, r = 1...M$ are the weights of the mixtures.

The product of the component Gaussian with its mixture weight i.e., $K_p(x_k)w_r$ is termed as component density. Sum of the component densities is given by Gaussian mixture density. The accuracy in capturing the true distribution of data depends on various parameters such as dimension of feature vectors, number of feature vectors and number of mixture components. In this work expectation maximization (EM) algorithm is used to train the GMM models using audio features.

3 Sentiment Analysis using Lyric Features

This section describes the process of extracting the textual lyrics of a song. These features are then used to build a classifier of positive or negative sentiment of a song. In Preprocessing step, lyrics which contain stanza names like "pallavi" and "charanam" were removed because, as the lyrics are collected from the Internet the headings ("pallavi" and "charanam") are common for each song which does not act like a feature to detect the sentiment of the song. If the same line has to repeated, it is represented as "x2" in the original lyrics, so "x2" is removed and the line opposite to that is considered as twice. For each song in a database one feature vector with 300 dimension is generated for better results. As we have 100 files, 100 feature vectors are generated one for each song. For checking the accuracy, each song is manually annotated and is given a tag like happy or sad.

Here Doc2Vec model is used for associating random documents with labels. Doc2vec modifies word2vec algorithm to a unsupervised learning of continuous representations for larger blocks of text such as sentences, paragraphs or whole documents means Doc2vec learns to correlate labels and words rather than words with other words. In the word2vec architecture, the two algorithms used are continuous bag of words and skip-gram and for the doc2vec architecture, the corresponding algorithms are distributed memory and distributed bag of words. All songs are given as input to the doc2vec. This generates a single vector that represents the meaning of a document, which can then be used as input to a supervised machine learning algorithm to associate documents with labels. Song sentiment analysis based on lyrics can be viewed as a text classification task which can be handle by SVM and NaiveBayes (NB) algorithms due to their better classification performance. Both SVM and NaiveBayes classifiers are trained with vectors generated from the doc2vec. After calculating the probabilities from both the classifiers, average probabilities of them is computed. Which ever class gives highest average probability that test case is hypothesized from that class. Like this these two classifiers are compared. By combining both the classifiers, rate of detecting the sentiment of a song is improved. Given a test data song, the trained models classifies it as either happy or sad. Three experiments are done on each song:beginning 30 seconds, last 30 seconds and for the whole song.

Table 1: Sentiment Classification with Lyric Features

	SVM	NB	SVM+NB
Whole song	60.6	52.3	70.2
Beginning of a song	67.5	57.3	**75.7**
Ending of a song	64.4	55.8	72.4

From Table 1 it is observed that a combination of both the classifiers gives high percentage for beginning of the song compared to the ending and whole song. Whole song gives less accuracy in detecting the sentiment of a song. By keeping the training data set constant several experiments are done on the test data. The average performance of sentiment analysis for beginning, ending and for whole song is 75.7, 72.4 and 70.2 respectively.

4 Sentiment Analysis using Audio Features

This section describes the process of extracting the audio features of a song. These features are then used to build a classifier of positive or negative sentiment of a song. Each song underwent the preprocessing step of converting mp3 files into wave file (.wav format), into 16 bit, 16000 Hz sampling frequency and to a mono channel. To extract a set of audio features like mfcc, chroma, prosody, temporal, spectrum, harmonics and tempo from a wave file openEAR/openSMILE toolkit [Eyben *et al.*, 2010] is used. Brief details about audio features are mentioned below:

- Prosody features include intensity, loudness and pitch that describe the speech signal.
- Temporal features also called as time domain features which are simple to extract like the energy of signal, zero crossing rate.
- Spectral features also called as frequency domain features which are extracted by converting the time domain into frequency domain using the Fourier Transform. It include features like fundamental frequency, spectral centroid, spectral flux, spectral roll-off, spectral kurtosis, spectral skewness. These features can be used to identify the notes, pitch, rhythm, and melody.
- In Mel-frequency Cepstral Coefficients (MFCC) (13 dimension feature vector) the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely.

- Chroma features (12 dimension feature vector) are most popular feature in music and is extensively used for chord, key recognition and segmentation.
- Harmonic tempo is the rate at which the chords change in the musical composition in relation to the rate of notes.

Although this toolkit is designed for the emotion recognition, the research has been done on sentimental analysis by using the same toolkit which is succeeded [Mairesse *et al.*, 2012]. As prosody have been used before for the task of emotion recognition in speech, it has also been experimented for the task of sentiment analysis by the authors [Mairesse *et al.*, 2012]. Three experiments are performed here:beginning 30 seconds, last 30 seconds and for the whole song. Features that are extracted are trained on the classifiers such as SVM, GMM and combination of both. GMM are known for capturing the distribution in the features and SVM are known for discriminating the features. Hence these models are combined improve the performance of detecting the sentiment of a song using the audio features. GMM need more features for training compared to Naive Bayes and SVM, but in textual part we have less features (only one feature vector for one song using doc2vec). Where as for audio, several features are their because for each song features are extracted at frame level with a frame size of 20 ms. So for acoustic models GMM and SVM are used where as for linguistic features Naive Bayes and SVM are used. A total of 40 dimension feature vectors are extracted, each of them obtained at frame level. During the feature extraction, frame size of 25ms and frame shift of 10ms are used. In this work, number of mixtures for GMM models (64) and Gaussian kernel parameters for SVM models are determined empirically.

Table 2: Sentiment Classification with Audio Features

	SVM	GMM	SVM+GMM
Whole song	52.8	54.9	69.7
Beginning of the song	55.8	73.5	**88.3**
Ending of the song	64.7	61.7	82.35

From Table 2 it is observed that the whole song gives less performance in detecting the sentiment of a song because the whole song will carries different attributes (happy and sad) which is not clear. So by using part of song, the performance is increased. Hence experiments are done even on beginning and ending of the song. Combination of both classifiers gives a high percentage for beginning of the song compared to the ending of the song. SVM is best performed at the ending of the song, GMM is best performed at the beginning of the song. By keeping training data set constant several experiments are done on the test data. The average performance of sentiment analysis for beginning, ending and for whole song is 88.3, 82.3 and 69.7 respectively.

5 Multimodal Sentiment Analysis

The main advantage that comes with the analysis of audio as compared to their textual data is it will have voice modularity. In textual data, the only source that we have is information regarding the words and their dependencies, which may sometime be insufficient to convey the exact sentiment of the song. Instead, audio data contain multiple modalities like acoustic, and linguistic streams. From our experiments it is observed that textual data gives less percentage than the audio, so the simultaneous use of these two modalities will help to create a better sentiment analysis model to detect whether the song is happy or sad.

Sequence of steps in proposed approach is presented in the Figure 1. Table 3 presents the accuracy of sentiment by combining lyrics and audio features. The whole song may not convey sentiment, so there will be lot of similarity between sad and happy features. Hence features extracted from different parts of a song are used to identify the sentiment of the song. To handle the similarity of sentiment classes, decision from different classification models trained using different modalities are combined. By combining both the modalities performance is improved by 3 to 5%.

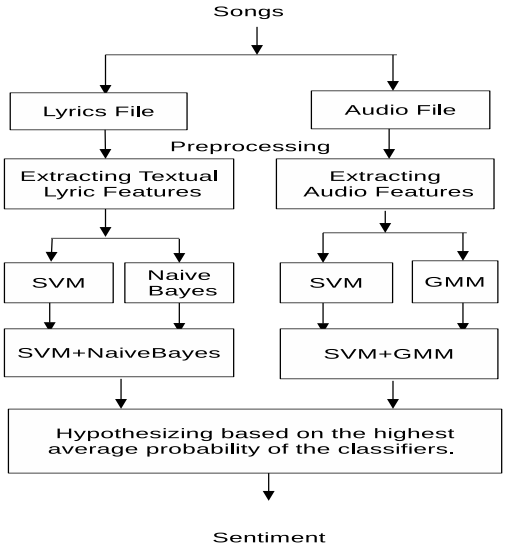

Figure 1: Block diagram of multimodal sentiment analysis of songs

Table 3: Sentiment Classification with Lyric and Audio Features

	Lyric	Audio	Lyric+Audio
Whole song	70.2	69.7	75.8
Beginning of a song	75.7	88.3	**91.2**
Ending of a song	72.4	82.3	85.6

6 Summary and Conclusions

In this paper, an approach to extract the sentiment of a song using both lyrics and audio information is demonstrated.

Lyric features which are generated using Doc2Vec and most efficient audio features like spectral, chroma, etc are used to built the classifiers. Sentiment analysis systems are built using the whole song, beginning of the song and ending of the song. By taking the whole song the performance is very less because the full song will contain more information (features) which is confusing. Hence experiments are done on the beginning and the ending of the songs which are giving better results. Features are extracted from beginning of the song are observed to be giving better performance compared to the whole song and the ending of the song. Because the instruments and vocals which convey the sentiment for beginning part of the song may or may not sustain throughout the song. Several experiments are done by keeping training data constant. The proposed method is evaluated using 100 songs. From the experimental results, recognition rate is observed to be in between 85% to 91.2%. This work can be extended by including more attributes like angry, fear and by extracting more features like rhythm and tonality. The percentage of lyric sentiment analysis can be improved by using rule based and linguistic approach.

References

[Campbel *et al.*, 2006] M William Campbel, P Joseph Campbell, A Douglas Reynolds, Elliot Singer, and A Pedro Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2):210–229, 2006.

[Dong *et al.*, 2010] Zhendong Dong, Qiang Dong, and Changling Hao. Hownet and its computation of meaning. In *Proc. 23rd international conference on computational linguistics: demonstrations, association for computational linguistic*, pages 53–56, 2010.

[Ellis, 2007] D. P. W. Ellis. Clasifying music audio with timbral and chroma features. In *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR)*, pages 339–340, 2007.

[Eyben *et al.*, 2010] F. Eyben, M. Wollmer, and B. Schulle. opensmile the munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, 2010.

[Hu and Downie, 2010] X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proc. Joint Conference on Digital Libraries, (JCDL)*, pages 159–168, 2010.

[Hu *et al.*, 2007] X. Hu, M. Bay, and J. S. Downie. Creating a simplified music mood classification ground-truth set. In *Proc. 8th International Conference on Music Information Retrieval*, 2007.

[Hu *et al.*, 2009] Xiao Hu, J. Stephen Downie, and Andreas F. Ehmann. Lyric text mining in music mood classification. In *Proc. 10th International Conference on Music Information Retrieval (ISMIR)*, pages 411–416, 2009.

[Jamdar *et al.*, 2015] Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence and Applications(IJAIA)*, 6(3):35–50, 2015.

[Kaushik *et al.*, 2013] Lakshmish Kaushik, Abhijeet Sangwan, and John H L. Hansen. Sentiment extraction from natural audio streams. In *proc. ICASSP*, pages 8485–8489, 2013.

[Kaushik *et al.*, 2015] Lakshmish Kaushik, Abhijeet Sangwan, and John H.L. Hansen. Automatic audio sentiment extraction using keyword spotting. In *Proc. INTERSPEECH*, pages 2709–2713, September 2015.

[Mairesse *et al.*, 2012] F. Mairesse, J. Polifroni, and G. Di Fabbrizio. Can prosody inform sentiment analysis? experiments on short spoken reviews. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 5093–5096, 2012.

[Medhat *et al.*, 2014] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering journal*, pages 1093–1113, 2014.

[Patra *et al.*, 2013] B. G. Patra, D. Das, and S. Bandyopadhyay. Unsupervised approach to hindi music mood classification. In *Mining Intelligence and Knowledge Exploration (MIKE 2013), R. Prasath and T. Kathirvalavakumar (Eds.):LNAI 8284*, pages 62–69, 2013. Springer International Publishing.

[Reynolds and Rose, 1995] A Douglas Reynolds and C Richard Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[Rish, 2001] Irina Rish. An empirical study of the naive bayes classifier. In *Proc. IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[Shanmugapriya and Dr.B.Srinivasan, 2015] K.P Shanmugapriya and Dr.B.Srinivasan. An efficient method for determining sentiment from song lyrics based on wordnet representation using hmm. *International Journal of Innovative Research in Computer and Communication Engineering*, 3(2):1139–1145, February 2015.

[Tyagi and Chandra, 2015] Atul Tyagi and Nidhi Chandra. An introduction to the world of sentimnt analysis. In *Proc. 28th IRF International Conference*, June 2015.

[Wang *et al.*, 2009] Tao Wang, DongJu Kim, KwangSeok Hong, and JehSeon Youn. Music information retrieval system using lyrics and melody information. In *proc. Asia-Pacific Conference on Information Processing*, pages 601–604, 2009.

[Xia *et al.*, 2008] Yunqing Xia, Linlin Wang, Kam-Fai Wong, and Mingxing Xu. Sentiment vector space model for lyric-based song sentiment classification. In *proc. ACL-08:HLT, Short Papers*, pages 133–136, 2008.

[Zhong *et al.*, 2012] Jiang Zhong, Yifeng Cheng, Siyuan Yang, and Luosheng Wen. Music sentiment classification integrating audio with lyrics. *Information and Computational Science*, 9(1):35–54, 2012.