ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

A New Technique for Opinion Mining of Hinglish Words

Subramaniam Seshadri[1], Ashwin Lohidasan[2], Rohini Lokhande[3], Ashwini Save[4], Tatwadarshi P.Nagarhalli[5]

Student, Dept. of Computer Engineering, VIVA Institute of Technology, Mumbai, India[1]

Student, Dept. of Computer Engineering, VIVA Institute of Technology, Mumbai, India [2]

Student, Dept. of Computer Engineering, VIVA Institute of Technology, Mumbai, India [3]

Assistant Professor, Dept. of Computer Engineering, VIVA Institute of Technology, Mumbai, India [4]

Assistant Professor, Dept. of Computer Engineering, VIVA Institute of Technology, Mumbai, India [5]

ABSTRACT: In recent years a lot of user generated content have been made available on the internet. This user generated content can be used as data sources in online systems. Opinions can be extracted from these user generated contents and be used for further analysis. This paper proposes a way, to build a classifier for detecting sentiments of Hinglish& English movie reviews. This has been implemented and tested to test the performance of the proposed technique as well.

KEYWORDS: Opinion mining, Naïve Bayes, sentiment analysis, machine learning, Hinglish, Hinglish analysis.

I. INTRODUCTION

The sentiments found within comments provide useful indicators which can be used for many purposes like pattern analysis and decision making. The sentiment can be classified as positive and negative and the overall sentiment of the comment can be calculated based on the number of sentimental keywords. Sentimental analysis facilitates policies to analyze public opinions with respect to particular subject.

A large volume of information is available in the textual form. This textual information can be either Opinions or Facts. The facts are the objective expressions which describe the entities, events and properties whereas the opinion is the subjective expression which describes people's opinions, emotions and sentiments towards entities and their properties [11].

Opinion mining (also known as Sentiment analysis) indicates the use of techniques such as text analysis, natural language processing, and computational linguistics to identify and extract subjective information [1]. It is used to analyze the attitude of the writer towards the commodity or service.

These large volumes of information that appear on the internet is very difficult to be processed by individuals manually, leading to information overload, hence affecting decision making process. In India, these reviews are often in Hinglish (Hindi words typed in English). Hinglish is a language used routinely in music, movies and it is through which millions of Indians communicate. As it used so widely, people tend to use it in their comments and reviews, to explicitly express their opinions. These comments and reviews when analyzed using a standard classifier have been found to be less effective when used for the Hinglish words. In order to overcome this obstacle, this paper proposes a technique of opinion mining where Hinglish words are also taken into account for producing better results.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

II. RELATED WORK

Related work in this field includes analyzing the data or information which is available in the textual format at different levels. They are as follows: [7]

Document level: At this level, the task is to identify whether a whole document expresses a positive or negative sentiment. It classifies a whole review as positive or negative. In this level of analysis it is assumed that the opinions expressed by a document are towards a single entity only.

Sentence level: At this level, the objective is to determine whether the opinion expressed by a sentence is positive, negative or neutral. This level of analysis relates to subjectivity classification which differentiates factual information from subjective views and opinions [2].

Entity and Aspect Level: In this level, analysis is done to identify towards which entity the opinion was directed. Also called feature level, it separates an opinion into sentiment and its intended target. An opinion without its target being identified is of limited use.

The document level analysis is of limited use as the opinions expressed by the document are towards a particular entity only. Thus, this approach is not applicable to comparative reviews.

The Sentence level is based on subjectivity. It differentiates facts from opinions. It is complex to implement compared to document level, but easier than Entity & Aspect Level.

The Entity & Aspect Level is the most complex analysis. It determines the entity towards which the opinion was directed. In most cases, classifiers built based on this analysis method are of little use.

III. PROPOSED METHOD

Generally the knowledge base which is used to for classification is the SentiWordnet. This SentiWordsnet contains ranking and ratings for only English words. But in India, to explicitly express ones opinion people use words which are not found in the English dictionary. These expressions by the people is often in Hinglish (Hindi words typed in English). Hinglish is a language used routinely by many people to express themselves while commenting or reviewing any music or movies or anything else.

For any online sentiment analysis a sentiment library is used, which is generally Sentiwordnet, which has only English words in the proposed system the sentiment library or the knowledge base contains, along with the English, Hinglish words for better accuracy. Here the system consists of a sentiment library i.e. training data consisting of positive and negative words. The system breaks user comments into tokens to check for sentiment keywords. The keywords, including both the English and Hinglish keywords, once found is associatedwith a sentiment rank to the comment. The system then amalgamates the ratings of both the English keywords as well as Hinglish key words and calculates an average score for that particular comment.This automated system calculates the average score of a comment after considering both the English and the Hinglish words, thus providing better results compared to other sentiment analysis system where Hinglish words are concerned.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

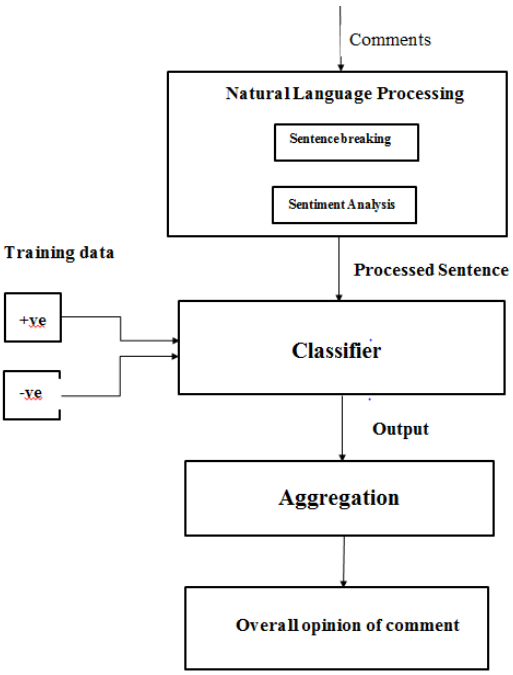

Fig 1: Flowchart of the System.

Figure 1 depicts a flowchart of the proposed system. The comment or review containing English and Hinglish words is fed as an input to the system. The comments are then broken into sentences (for multi-sentence comment). These sentences are further divided into tokens. These tokens are then fed to the classifier. The classifier checks for Positive & Negative English and Hinglish words and assigns a word count to the particular class (positive or negative). Depending upon these words, the overall sentiment of the sentence is calculated. The same is done for each sentence. It is then aggregated to get the overall opinion of the comment.

The working of the system can be explained with an example as well.

“Bole tohmaza ah gaya”. “Full paisa wasool, zarabhipakaonahihai”. “Movie is a visual masterpiece”. “There were some shots, which made audience go crazy”. “Some scenes were as epic as 2001: A Space Odyssey. Matthew McConaughey does what he is best in and supposed to”. “Go, sit back and enjoy. Also, make sure you go to the best theatre close to you”. “If you loved Donnie Darko, 2001: A Space Odyssey, this is the one for you, Get ready to get mindblown!!” “Interstellar Movie is EPIC in all sense!!” [13].

The above shows some examples of a movie reviews which contains Hinglish comments. After receiving this comments as an input the system now breaks the user comments to find out keywords that help to find out the sentiment.

Bole tohmaza ah gaya. Full paisa wasool, zarabhipakaonahihai. Movie is a visual masterpiece. There were some shots, which made audience go crazy. Some scenes were as epic as 2001: A Space Odyssey. Matthew McConaughey does what he is best in and supposed to. Go, sit back and enjoy. Also, make sure you go to the best theatre close to you. If you loved Donnie Darko, 2001: A Space Odyssey, this is the one for you, Get ready to get mindblown!! Interstellar Movie is EPIC in all sense!!

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

The words which are underlined in the above table are those words which give the sentiment of the sentence. Now these words are checked with the training data to identify the polarity of the sentence. The training data consist of set of positive words and a set of negative words.

Table 1. Assigning polarity based on training data

Maza	1
Wasool	1
Pakao	-1
masterpiece	1
Epic	1
Best	1
enjoy	1
loved	1
mindblown	1

Table 1 gives an example as to how the polarities have been assigned to the words in the above example which contains Hinglish words. The Hinglish words are also assigned polarities, this gives a better or more accurate sentiment of the sentence.

As we can find the number of positive words is more than the number of negative words, hence the above comment is considered to be a positive comment.

The system does same for all the comments of a particular movie and generates an overall rating for the movie. This score is generated for all the movies in the system.

IV. RESULTS AND ANALYSIS

The Naive Bayes Classifier gives an approximate accuracy of 0.8 which is greater than most classifiers. Note that this accuracy is only for English comments. So, when Hinglish words are fed to this classifier it accuracy decreases. In the proposed system the classifier is for the classification of the Hinglishwords along with the English comments as well, hence the accuracy provided by this system is higher when the comments include Hinglish words. This is due to the fact that, the proposed system will classify English as well as Hinglish comments, which will be ignored by a standard classifier.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

The performance can be well analyzed with the help of an example given below.

Fig 2: Standard Classifier

Fig 3: System which can process Hinglish words

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

Figure 2 depicts the classification of a Hinglish comment entered by the user. Figure 3 depicts the classification of the same comment by our system. In the Figure 2, the user comment "pakao movie hai" (It was a boring movie) is processed by the standard classifier. Since, this classifier contains only positive & negative English sentences as its training data, it won't be able to process the said comment and give appropriate rating. As seen in figure 2,the classifier generates a rating of '0' for the said input comment. In other cases, it may generate inappropriate or garbage rating, since it can't recognize Hinglish words.

In the figure 3, the same user comment when used as input in our system, generates a rating of '1'. This is due to the fact that the input comment which contains the Hinglish words are also considered for classification and sentiment analysis. Due to this, the Hinglish words which get ignored during the classifier stage get acknowledged with their proper rating in the processing.

V. CONCLUSION

This paper provides a method in which sentimental analysis for Hinglish comments can be done. The system breaks user comments to check for sentiment. The classifier has been trained with positive and negative words. Once it analyzes the sentence, it associates a sentiment rank (positive or negative) with the comment. This system does sentimental analysis for Hinglish comments as well hence its accuracy is greater than other systems where Hinglish comments are involved. Sentiment analysis can be used in different fields when it is required for identifying the task of whether the opinion expressed in a text is positive or negative in general, or about a given topic. The important thing is that the system is robust. Also provision is provided for future developments in the system. In future, more work can be done to further improve the performance.

REFERENCES

[1] Turney, P.,"Proceedings of the Association for Computational Linguistics", pp. 417–424, 2002.
[2] SmeureanuI., and Bucur C., "Applying Supervised Opinion Mining Techniques on Online User Reviews", InformaticaEconomica, 2012.
[3] Liu, B.,"Sentiment Analysis & Opinion Mining", 2012.
[4] Jagtap, V. S., and Pawar, K.,"Analysis of different approaches to Sentence-Level Sentiment Classification", International Journal of Scientific Engineering and Technology, pp. 164-1702013.
[5] ThelwallM., and Prabowol, R., "Sentiment Analysis: A Combined Approach" http://www.cyberemotions.eu/rudy-sentiment-preprint.pdf.
[6] Rish, I.,"An empirical study of the Naïve Bayes Classifier", T. J. Watson Research Center, November 2001.
[7] Liu, B.,"Opinion Mining and Sentiment Analysis", AAAI, San Francisco, 2011.
[8] Chaovalit, Zhou, L.,"Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
[9] Shelke, N., Deshpande S., and Thakre, V.,"Survey of Techniques for Opinion Mining",International Journal of Computer Applications (0975 – 8887)Volume 57– No.13, November 2012
[10] The Art of Opinion Mining and Its Application Domains: -A Survey (ICRTITCS - 2012)
[11] Mishra N., and JhaC. K.,"Classification of Opinion Mining Techniques", International Journal of Computer Applications (0975 – 8887)Volume 56– No.13, October 2012
[12] VinodhiniG., and Chandrasekaran, R. M.,"Sentiment analysis and Opinion Mining: A survey", International Journal of advanced Research in Computer Science and Software Engineering, Vol. 2 Issue 6, 2012.
[13] http://www.imdb.com/title/tt0816692/reviews, last accessed: 10th June 2015, 11.01 AM.