SOCIAL MEDIA INTERACTION FOR DETECTING FAKE REVIEWS

V.Parameshwarreddy[1],S.Ashokkumar[2],B.PanduRangaRaju[3]

[1]Assistant Professor, Dept of IT,Annamacharya institute of technology and sciences, Rajampet,

[2]Assistant Professor, Dept of IT,Annamacharya institute of technology and sciences, Rajampet,

[3] Assistant Professor, Dept of IT, Annamacharya institute of technology and sciences, Rajampet.

Abstract- Social media monitoring has been growing day by day so analyzing of social data plays an important role in knowing customer behavior. So we are analyzing Social data such as Twitter Tweets using sentiment analysis which checks the attitude of User review on movies. This paper develops a combined dictionary based on social media keywords and online review and also find hidden relationship pattern from these keyword. In recent years, shopping online is becoming more and more popular. When it need to decide whether to purchase a product or not on line, the opinions of others become important. It presents a great opportunity to share our viewpoints for various products purchase. However, people face the information overloading problem. How to mine valuable information from reviews to understand a user's preferences and make an accurate recommendation is crucial. Traditional recommender systems consider some factors, such as user's purchase records, product category, and geographic location. In this work, it propose a sentiment-based rating prediction method to improve prediction accuracy in recommender systems. Firstly, it propose a social user sentimental measurement approach and calculate each user's sentiment on items. Secondly, it not only consider a user's own sentimental attributes but also take interpersonal sentimental influence into consideration. Then, consider item reputation, which can be inferred by the sentimental distributions of a user set that reflect customers' comprehensive evaluation. At last, by fusing three factors-user sentiment similarity, interpersonal sentimental influence, and item's reputation similarity into recommender system to make an accurate rating prediction. It conduct a performance evaluation of the three sentimental factors on a real-world dataset. Experimental results show the sentiment can well characterize user preferences, which help to improve the recommendation performance.

Keywords: e-commerce, product recommender, product demographic, microblogs, recurrent neural networks.

1. INTRODUCTION

Nowadays, Social media is becoming more and more popular since mobile devices can access social network easily from anywhere. Therefore, Social media is becoming an important topic for research in many fields. As number of people using social network are growing day by day, to communicate with their peers so that they can share their personal feeling everyday and views are created on large scale. Social Media Monitoring or tracking is most important topic in today's current scenario. In today many companies have been using Social Media Marketing to advertise their products or brands, so it becomes essential for them that they can be able to calculate the success and usefulness of each product [2]. For Constructing a Social Media Monitoring, various tool has been required which involves two components: one to evaluate how many user of their brand are attracted due to their promotion and second to find out what people thinks about the particular brand. To evaluate the opinion of the users is not as easy as it seems to all users. For evaluating their attitude may requires to perform Sentiment Analysis, which is defined as to identify the polarity of customer behavior, the subjective and the emotions of particular document or sentence. To process this we need Machine Learning and Natural Language Processing methods and this is place where most of the developers facing difficulty when they are trying to form their own tools. Over the recent years, an emerging interest has been occurred in supporting social media analysis for advertising, opinion analysis and understanding community cohesion. Social media data adapts to many of the classifications attributed for "big-data" – i.e. volume, velocity and variety. Analysis of Social media needs to be undertaken over large volumes of data in an efficient and timely manner. Analysing the media content has been centralized in social sciences, due to the key role that the social media plays in modelling public opinion. This type of analysis typically on the preliminary coding of the text being examined, a step that involves reading and annotating the text and that limits the sizes of the data that can be analysed. With the development of Web, more and more people are connecting to the Internet and becoming information producers instead of only information consumers in the past, resulting to the serious problem, information overloading. There is much personal information in online textual reviews, which plays a very important role on decision processes. For example, the customer will decide what to buy if he or she sees valuable reviews posted by others, especially user"s trusted friend. People believe reviews and reviewers will do help to the rating prediction based on the idea that high-star ratings may greatly be attached with good reviews. Hence, how to mine reviews and the relation between reviewers in social networks has become an important issue in web mining, machine learning and natural language processing. It focus on the rating prediction task.

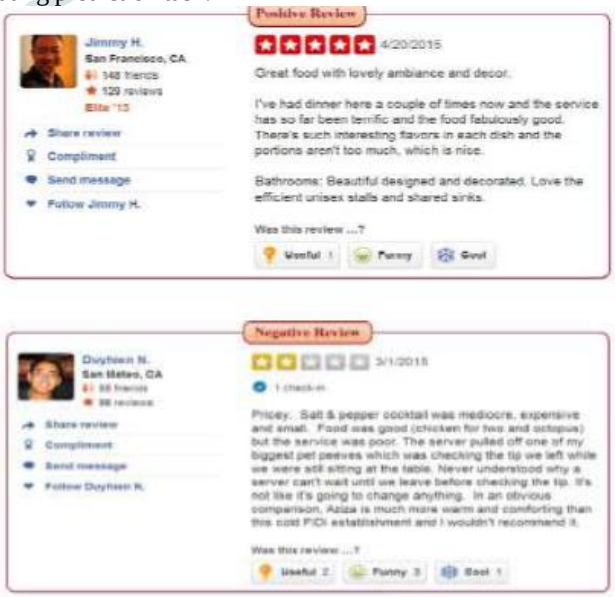

Fig 1: An Example of Positive Review and Negative review on websites

In Fig.1, we intuitively show an example of positive reviews and negative reviews on website. From Fig.1, there are many positive words in a 5-star review, such as “great”, and “lovely”. But in a 2-star review we find negative words, such as “expensive”, and “poor”. That means a good review reflects a high star-level and a bad review reflects a low-level. When we know the advantages and disadvantages from the two kinds of reviews, we can easily make a decision.

Sentiment Analysis

Sentiment analysis refers to the use of natural language processing to identify and extract one-sided information in source materials or simply it refers to the process of detecting the polarity of the text. It also referred as opinion mining, as it derives the opinion, or the attitude of a user. A common approach of using this is described how people think about a particular topic. Sentiment analysis helps in determining the thoughts of a speaker or a writer with respect to some subject matter or the overall contextual polarity of a document. The attitude may be his or her decision or estimate, the emotional state of the user while writing.

Sentiment Analysis is hard

Today, Sentiment analysis plays an important role where various machine learning technique is used in determining the sentiment of very huge amounts of text or speech. Various application tasks include such as determining how someone is excited for an upcoming movie, correlates different views for a political party with people’s positive attitude towards vote for that party, or by converting written hotel reviews into 5-star based on scaling across categories like ‘quality of food’, ‘services’, ‘living room’ and ‘facilities’ provided. As there is huge amount of information is shared on social media, forums, blogs, newspaper etc. it is easy to see why there is a need for sentiment analysis as there is much information to process manually which is not possible in today’s time.

Text Analysis process

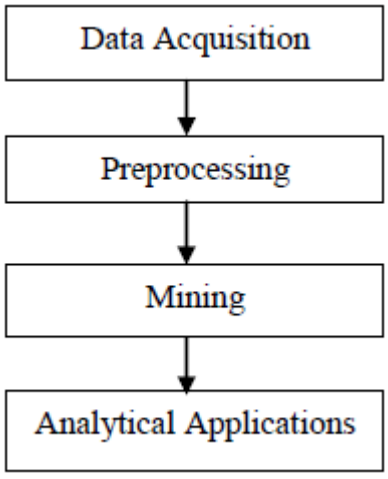

Fig 2: Process of Analyzing Text

This process involves the following steps [23]:

Data Acquisition: In this data acquisition, data are gathered from different relevant sources such as web crawling, twitter tweets, online review, newsfeeds, document scanning etc.

Preprocessing: It is used to remove noisy, inconsistent and incomplete data. For doing the classification, Text preprocessing and feature extraction is a preliminary phase.

Preprocessing involves 3 steps:

Tokenization or segmentation: It is the process of splitting a string of written language into its words. Text data consists of block of characters referred to as tokens. So the documents are being separated as tokens and have been used for further processing.

Removal of stop words: Stop words are the words which are needed to be filtered i.e. may be before or after natural language processing. Stop words are words which contain little informational. Various tools specifically avoid to remove these stop words in order to support phrase search. Several collections of words can be chosen as stop words for any purpose. Some search engines, removes most of the common words which include lexical words such as "want" from a text in order to improve performance. Search engine or natural language processing may contain a variety of stop words. It includes English stop words such as “and”, “the”, “a”, “it”, “you”, “may”, “that”, “I”, “an”, “of” etc. which are considered as ‘functional words’ as they don’t have meaning.

Researchers have shown that by removing stop words from the file, you can get the benefit of reduced index size without much affecting the accuracy of a user's. But care should be taken however to take into consideration the user's needs. Mostly, all search engines helps in eliminating the stop words from their indexes. With the help of eliminating stop words from the index, the index size can be reduced to about 33% for a word level index. While assessing the content of natural language processing, meaning of word can be conveyed more clearly by removing the functional word [21]. Stemming: It is the term which used to describe the process to reduce derived words to their origin word stem. Since 1960s, algorithms for stemming have been studied in the field of computer science. Different Stemming methods are commonly referred as stemming algorithms or stemmers. For English, the stemmer example are that, it should identify the string “cats”, ”catty” as based on the root word “cat”, and also “walks”, “walked”, “walking” as based on the root word "walk" [22].

Data Mining: Applying different mining techniques to derive usefulness about stored information. Different mining approaches are classification, clustering, statistical analysis, natural language processing etc. In text analytics, mainly classification technique is used. Classification is a supervised learning method that helps in assigning a class label to an unclassified tuple according to an already classified instance set. Data classifying and identifying is all about to tag the data so it can be create quickly and efficiently. But various organizations can gain from re-transforming their information, which helps in order to cut storage and backup costs, with increasing the speed of data searches. Classification can help an organization to meet authorized and regulatory requirements to retrieve specific information within a specific time period, and this is most important factor behind implementing various data classification technology. Analytical Application: It provides valuable things from text mining so that it can provide information that helps in improving decision and processes. It includes following ways such as sentiment analysis, document imaging, fraud analysis etc.

II LITERATURE SURVEY

This paper chose primarily three methods for text classification because of their relative popularity and success in prediction of sentiments:

- **Naive Bayes:** This works on the assumption of conditional independence and despite this oversimplified assumption, Naive Bayes performs well in many complex real-world problems. Naive Bayes classifier is superior in terms of CPU and memory consumption.
- **Support Vector Machines:** SVM also provides a robust approach to build text classifiers and was picked because of its ability to handle High dimensional input space. When learning text classifiers, many (morethan10000) features can been countered. Since SVMs use over fitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces.
- **Maximum Entropy:** MaxEnt Naïve Bayes is based on conditional independence assumption, hence to ensure that this paper covers an alternative, it uses Maximum Entropy that does not assume conditional independence. It is based on the Principle of Maximum Entropy and from all the models that fit the training data, selects the one which has the largest entropy. Although it takes more time than Naïve Bayes to train the model, this method has proven to be useful in cases where we do not know anything about the prior distribution

(Hening-Thurau etal., 2003) state that customer comments articulated via the Internet are available to a large number of other customer's, and therefore can be expected to have a significant impact on the success of goods and services. This on consumer buying and communication behavior are tested in a large-scale empirical study. The results illustrate that consumers read online articulations mainly to save decision-making time and make better buying decisions. Structural equation modeling shows that their motives for retrieving online articulations strongly influence their behavior (Duan et al., 2008) showed that both a movie's box office revenue and WOM valence significantly influence WOM volume. WOM volume in turn leads to higher retrieve other customer's online articulations from webbased consumer opinion platforms. The relevance of these motives and their impact box office performance. This positive feedback mechanism highlights the importance of WOM in generating and sustaining retail revenue. (Chevalier & Mayzlin, 2006) hypothesized that buyers suspect that many reviewers are authors or other biased parties. They found marginal (negative) impact of 1-star reviews is greater than the (positive) impact of 5-star reviews. The results suggest that new forms of customer communication on the Internet have an important impact on customer behavior. Work on sentiment analysis found using a formal approach is the work by (Simancík and Lee, 2009). The paper presents a method to detect sentiment of newspaper headlines, in fact partially using the same grammar formalism that later will be presented and used in this work, however without the combinatorial logic approach. The paper focus on some specific problems arising with analysing newspaper headlines, e.g. such as headline texts often do not constitute a complete sentence, etc. However the paper also present more general methods, including a method for building a highly covering map from words to polarities based on a small set of positive and negative seed words. This method has been adopted by this thesis, as it solves the assignment of polarity values on the lexical level quite elegantly, and is very loosely coupled to the domain. However, their actual semantic analysis, which unfortunately is described somewhat shallow in the paper, seems to suffer from severe problems with respect to certain phrase structures, e.g. dependent clauses. eWOM is a form of communication, defined as a: "statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet" (Hennig-Thurau, Gwinner, Walsh , & Gremle, 2004,p. 39). eWOM may be less personal in that it is not face-to-face (or maybe just personal in a different way than in the past), but it is more powerful because it is immediate, has a significant reach, is credible by being in print, and is accessible by others (Hennig-Thurau et al., 2004.

Collaborative Filtering

Collaborative filtering (CF) is an important and popular technology for recommender systems. The task of CF is to predict user preferences for the unrated items, after which a list of most preferred items can be recommended to users. The methods are classified into user-based CF and item-based CF. The basic idea of user-based CF approach is to find out a set of users who have similar favour patterns to a given user (i.e., „neighbours" of the user) and recommend to the user those items that other users in the same set like, while the item-based CF approach aims to provide a user with the recommendation on an item based on the other items with high correlations (i.e., „neighbours" of the item). In all collaborative filtering methods, it is a significant step to find users" (or items") neighbours, that is, a set of similar users (or items). Currently, almost all CF methods measure users" similarity (or items" similarity) based on co-rated items of users (or common users of items). Collaborative filtering and content based filtering have been widely used to help users find out the most valuable information.

Matrix Factorization based Approaches

1) Basic Matrix Factorization

Matrix factorization is one of the most popular approaches for low-dimensional matrix decomposition. Matrix factorization based techniques have proven to be efficient in recommender systems when predicting user preferences from known user-item ratings. Matrix can be inferred by decomposing item reviews that users gave to the items. Matrix factorization methods have been proposed for social recommendation due to their efficiency to dealing with large datasets. several matrix factorization methods have been proposed for collaborative filtering. The matrix approximations all focus on representing the user-item rating matrix with low-dimensional latent vectors.

2) Social Recommendation

In real life, people"s decision is often affected by friends" action or recommendation. How to utilize social information has been extensively studied. Yang et al. [6] propose the concept of "Trust Circles" in social network based on probabilistic matrix factorization. Jiang *et al.* [7] propose another important factor, the individual preference. some websites do not always offer structured information, and all of these methods do not leverage users" unstructured information, i.e. reviews, explicit social networks information is not always available and it is difficult to provide a good prediction for each user. For this problem the sentiment factor term is used to improve social recommendation.

III. EXISTING SYSTEM

Different types of data are generated from different Social media groups that need to be organized and to monitor people's attitude towards products, gadgets, movie review etc. This database is collected from different social media sites for example Twitter, Face book, Online review, shopping sites etc. Text analytics and Sentiment analysis can help to develop valuable business insights from text based contents that may be in the form of word documents, tweets, comments and news that related to Social media. The foremost reason of Sentiment analysis is so complex is that words often take different meanings and are associated with different emotions depending on the domain in which they are being used. Dataset is analyzed by using the weka tool. The hidden relationship has to be extracted from this type of database using different mining approaches in Weka tool. Dictionary building for detailed sentiment analysis implies making an initial list of adjectives and nouns which are normally used when describing a specific movie review. Phrases and terms are extracted from this relational dataset and their meaning has been added to dictionary for next generation analysis. In tweets, informal and shortcuts has been used for explaining terms or views and this is done with the help of sentiments analysis is not an easy process. To reduce this, data mining approaches has been used for extraction of features from these datasets.

IV. PROPOSED WORK

The proposed method comprises main components: Identify social relation between users, sentiment dictionaries, Recommendation system and User. The purpose of approach is to find effective clues from reviews and predict social users" ratings. We firstly extract product features from user review corpus, and then we introduce the method of identifying social users" sentiment At last we fuse all of them into our sentiment-based rating prediction method. It proposed a Highest rating recommendation system for products and items. The contributions can be summarized as follows: It propose a recommendation system for food items. To develop the recommendation system, rating data sets of products and items in the particular category which is used to read the textual reviews given by the users. The main categories which are used in the application are nothing but Lectures & books, Fashions, Food & Drink, Sports, Kids & Family, Electronic appliances. The datasets used in this recommendation system are "DouBan" and "Yelp" and other review websites provides a broad thought in mining user preferences and prediction users ratings. And other dataset used is nothing but "Online Product Rating"

Textual reviews obtained from data sets is categorised into three types: To identify positive reviews, To identify negative reviews and To identify neutral reviews. With the help of these types of reviews we can identify the social relation between users which will help to categories the item. Fig 3 shows how review analysis is done form the original reviews on the websites. Sentimental dictionaries will give the information of brands, quality and price on the basis of matrix
factorization. This matrix factorization can be performed by using two types of methods which are by applying conjunctive rules and another is by comparing product feature and sentiment words. This matrix factorization method will ultimately give the highest rating product recommendation for all types of products and items to the user. This recommendation system can be used by the user to select which items to be ordered or purchased and which are not. This recommendation system will help to take any decisions for any type of product.

Fig 3: Review Analysis

V. METHODOLOGY

In this proposed system, hadoop open source data mining tool has been used so as to perform sentiment classification on movie review dataset. Here, goal is to classify dataset into positive and negative and form the combined dictionary of Twitter dataset and online review dataset. Main steps are:

Generating Dataset

Two dataset were collected firstly, from Twitter tweets and secondly, from Online review Dataset. The online review dataset consists of around 800 user's review archived on the IMDB (Internet Movie Database) portal. And for, Twitter dataset around 1000 review were collected and each review were formatted according to .arff file where review text and class label are only two attributes. Class label represent the overall user opinion. Here, we set simple rules for scaling the user review. For dataset, a user rating greater than 6 is considered as positive, between 4 to 6 considered as neutral and less than 4 considered as negative.

Preprocessing

For doing the classification, Text preprocessing and feature extraction is a preliminary phase. Preprocessing involves 3 steps: I. Word parsing and tokenization: In this phase, each user review splits into words of any natural processing language. As movie review contains block of character which are referred to as token.
II. Removal of stop words: Stop words are the words that contain little information so needed to be removed. As by removing them, performance increases. Here, we made a list of around 320 words and created a text file for it. So, at the time of preprocessing we have concluded this stop word so all the words are removed from our dataset i.e. filtered.

III. Stemming: It is defined as a process to reduce the derived words to their original word stem. For example, "talked", "talking", "talks" as based on the root word "talk". We have used Snowball stemmer to reduce the derived word to their origin.

Classification

Classification is a supervised learning method that helps in assigning a class label to an unclassified tuple according to an already classified instance set. Here, naïve bayes multinomial classifier has been used. Quality measure will be considered on the basis of percentage of correctly classified instances. For the validation phase, we use 10- fold cross validation method. Naïve bayes multinomial helps in generating dictionary and frequent set. It counts the occurrences of words in whole dataset and forms a dictionary of some most frequently occurring words.

The online review dataset consists of around 800 user's review archived on the IMDB (Internet Movie Database) portal. And for, Twitter dataset around 1000 review were collected and each review were formatted according to .arff file where review text and class

label are only two attributes. Here, we analyse the dataset based on accuracy given by naïve bayes multinomial. Online review dataset accuracy around 94.968% and for twitter its around 82.695%. Results show that we get better accuracy for online review as compared to twitter tweets as online review are more clear and in detail compare to twitter tweets.

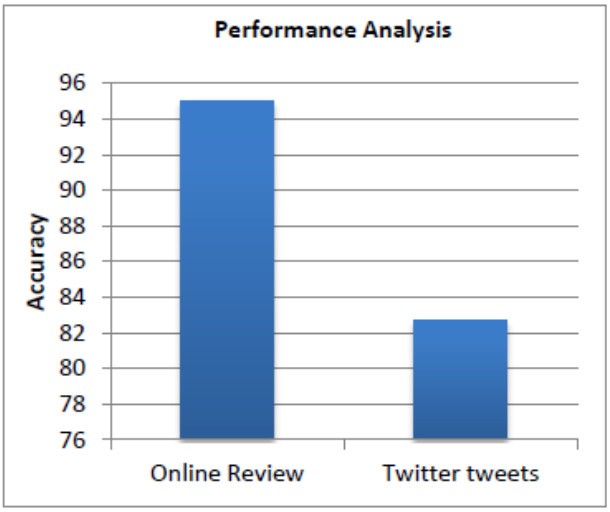

Fig 4: Performance analysis based on accuracy

Combined dictionary of words of twitter tweets and online review are formed based on probability of each word as we get by classification algorithm i.e. naïve bayes multinomial.

VI. CONCLUSION

Social media Monitoring has been growing very rapidly so there is a need for various organizations to analyze customer behavior or attitude of particular product or any movie review. So, the concepts of sentiment analysis have been introduced. Text analytics and sentiment analysis can help organization to derive valuable business insights. Attitude can be calculated based on polarity check. Sentiment analysis on Online review are done by forming dictionary which shows that it is easier to build dictionary on phrases but complex in case of Twitter as tweets consist of short hands as online review were written in more clear way as compared to Tweets. So, form hidden relationship between different keywords and a dictionary of the words on the basis of different categories of comments & tweets. Future work include to determine their features for the movie in detail i.e. make polarity check on different features such as actors, directors, scripts, music etc. and make the dictionary for them.

REFERENCE

[1] B. Wang, Y. Min, Y. Huang, X. Li, F. Wu, " Review rating prediction based on the content and weighting strong social relation of reviewers," in *Proceedings of the 2013 international workshop of Mining unstructured big data using natural language processing, ACM*. 2013, pp. 23-30.

[2] D. Tang, Q. Bing, T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proc. 53th Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, July 26-31, 2015, pp. 1014–1023.

[3] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014.

[4] W. Zhang, G. Ding, L. Chen, C. Li , and C. Zhang, " Generating virtual ratings from Chinese reviews to augment online recommendations," *ACM TIST*, vol.4, no.1. 2013, pp. 1-17.

[5] X. Lei, and X. Qian, "Rating prediction via exploring service reputation," *2015 IEEE 17th International Workshop on Multimedia*

[6] X. Yang, H. Steck, and Y. Liu, "Circle-based recommendation in online social networks, " in *Proc. 18th ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, Aug. 2012, pp. 1267–1275.

[7] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang, "Social contextual recommendation," in proc. *21st ACM Int. CIKM*, 2012, pp. 45-54.

[8] Z. Fu, X. Sun, Q. Liu, et al., "Achieving Efficient Cloud Search Services: Multi-Keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing," *IEICE Transactions on Communications*, 2015, 98(1):190-200.

[9] Y. Ren, J. Shen, J. Wang, J. Han, and S. Lee, "Mutual Verifiable Provable Data Auditing in Public Cloud Storage," *Journal of Internet Technology*, vol. 16, no. 2, 2015, pp. 317-323.

[10] W. Luo, F. Zhuang, X. Cheng, Q. H, Z. Shi, "Ratable aspects over sentiments: predicting ratings for unrated reviews," *IEEE International Conference on Data Mining (ICDM)*, 2014, pp. 380-389.

[11] T. Nakagawa, K. Inui, and S. Kurohashi, "Dependency tree-based sentiment classification using CRFs with Hidden Variables," *NAACL*, 2010, pp.786-794.

[12] Xiaojiang Lei, Xueming Qian, *Member, IEEE,* and Guoshuai Zhao, "Rating Prediction based on Social Sentiment from Textual Reviews," *IEEE Transactions On Multimedia*, MANUSCRIPT ID: MM-006446

[13] Mrs. R.Nithya, Dr. D.Maheshwari. 2014 Sentiment Analysis on Unstructured Review, International Conference on Intelligent Computing Application, IEEE.

[14] Ms. K. Mouthami, Ms. K.Nirmala Devi, Dr. V.Murali Bhaskaran. 2010 Sentiment Analysis and Classification based on Textual Reviews, Dept of CSE, Tamil Nadu, IEEE.

[15] Nargiza Bekmamedova, Graeme Shanks 2013 Social Media Analytics and Business Value: A Theoretical Framework and Case Study, Department of Computing and Information Systems, University of Melbourne.

[16] Simona Vinerean, Iuliana Cetina 2013 The Effects of Social Media Marketing on Online Consumer Behavior, International Journal of Business and Management; Vol. 8, No. 14.

[17] SitaramAsur, Bernardo A.Huberma 2012 Predicting the Future with Social Media, Social Computing Lab, HP Labs, Palo Alto, California.

[18] V.K. Singh, R.Piryani, A. Uddin, P.Waila. 2013 Sentiment Analysis of Movie Reviews, Department of Computer Science, New Delhi, India, Published in IEEE.

[19] V. S. Jagtap, Karishma Pawar. 2013 Analysis of different approaches to Sentence-Level Sentiment Classification, published in IJSET.

[20] Ya-Ting Chang, Shih-Wei Sun 2013 A Real time Interactive Visualization System for Knowledge Transfer from Social Media in a Big Data, Center for Art and Technology, Taipei National University of the Arts, Taipei, Taiwan, IEEE.

[21] Stop words: http://en.wikipedia.org/wiki/Stop_words

[22] Stemming: http://en.wikipedia.org/wiki/Stemming

[23] Decideo Amérique Du Nord Inc www.decideo.fr/bruley/docs/2sentiment_a_v0.ppt