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. 5, Issue 6, June 2016

Web Log Mining Operations for Classification of Product Reviews

Richa Pandey[1], Minu Choudhary[2]

P.G. Student, Department of Computer Science, Rungta College of Engineering & Technology, Bhilai, C.G., India[1]

Reader, Department of Computer Science, Rungta College of Engineering & Technology, Bhilai, C.G. , India[2]

ABSTRACT: Today is the era of internet. People directly or indirectly associated with the World Wide Web. Most of the works from small electric bill payment to bank transactions, from shopping to big business deals all can be easily done via internet. It is treated as a new source of entertainment, communication, education, etc. Apart from using internet people also take interest in sharing their personal suggestions and feedback in various micro-blogging sites (such as twitter), blogs, forums, social networks, etc. These reviews are very useful for the buyers in making decision as well as for the manufacturers or retailers to improve their quality product or services. As a result, a wide collection of consumer reviews are available on the net. It is a tedious task to go through each and every review before any purchase, so here arises the need of such system which helps in information retrieval. Also one major problem is fake reviewer problem or opinion spamming in the system, which highly affect the marketing of the product. In this paper we propose a product ranking framework which rank the product according to their quality and also filter the spam reviews by adopting the web log mining techniques.

KEYWORDS: Social media, Text mining, Opinion mining, Web log mining, Product reviews.

I. INTRODUCTION

The best tool to connect with consumer and industry leaders is the Social Media. With the arrival of social media the whole concept of marketing has been changed. New ways were opened for both consumers and marketers that before hardly exist or were too complicated. Now we can share a piece of content all over the world at our finger tips within a seconds. Many of us use social media for our personal reasons but now it's a necessity for the marketers to enter into social media to boost their business. For any business communication plays a key role. A positive communication can help to improve the reputation of any business. Social marketing will get more sales.

A new concept of reviews and rating system is emerging in the field of marketing. It is the most important and powerful tool nowadays because of the viral nature of social marketing.Online review reflects the strength and weakness of the business. It is helpful not only for the marketers but also play a vital role in consumer's decision making. Review and rating system provide an open channel for the people to post their comments and rank the product or services. Today's scenario totally depend on online reviews before making any purchase. Many research of marketing prove customer's reviews as a sales driver. There are mainly 3 categories of review: Positive review, Negative review and Neutral review. These reviews helps in eliminating any doubts (if any customer have) about any product or services and also help in product selection. Positive reviews builds confidence in user and their decision. On the other hand, negative reviews are also valuable for the manufactures to understand where they are lacking and do improvements.

Review and Rating System also called as Recommendation System. These system do suffer from some problems: Sparsity problem, Early-rate problem, Lack-of-negative-rating problem, Fake Reviewer problem.

In our paper, we are dealing with Fake Reviewer Problem. Fake reviewer problem arises when people or group of people post fake review to any product or service. To gain profit or to demote any target product or service, people cheat the system by opinion spamming.

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. 5, Issue 6, June 2016

The main aim of our project is to classify the products in terms of their quality (very good, good, average, poor, very poor) which is a prediction from the users review about the product and also to eliminate the fake reviewer problem by using Web Usage Mining operations.The organization of this paper is as follows:
Section 2: explains the basic concept. In this section we give overview of terms and terminology use in the project.
Section 3 describes our Proposed Methodology. Here we elaborate our algorithm.
Section 4 contains the Results and Discussion section. In this section we graphical represent the outcome of the proposed algorithm
Section 5 conclude our project. The section highlights our project and mention some works which should be carried out in future.

II. BASIC CONCEPTS

2.1 Text Mining

Text Mining also termed as *Text Data Mining or Text Analytics*, refers to process of extracting relevant information from an unstructured text document.Text Mining usually involves: gathering unstructured or raw information (from websites, emails, documents etc.), convert into a structured format (the pre-processing) and finally apply data mining techniques (clustering, classification, etc.) to derive high-quality information. Text mining also covers other fields of data mining, web mining, information retrieval, natural language processing (NLP).

2.2 Opinion Mining

Opinion Mining can be called as sentiment analysis, which is a process of detecting public's mood towards an entity. It is a type of natural language processing.
Opinion mining plays a vital role in various field of business and marketing. It helps marketers to evaluate their success or failure in business. It also helps customers in decision making before any purchase or use of any service. Basically sentiment analysis or opinion mining is the study of human behaviour or mood towards an entity.

There are 3 levels of sentiment analysis (SA) classification:

- *Document -level SA*
- *Sentence-level SA*
- *Aspect-level SA*

Document-level SA consider the whole document and classify the sentiment as positive or negative. *Sentence-level SA* deals with each sentence within a document and extract opinion at each sentence. It is true to say sentence as a short documents. The problem with document-level and sentence-level SA is that they does not provide relevant opinion on all aspect of the entity. To overcome this, one should go for Aspect-level SA. *Aspect-level SA* aims to find opinion with respect to each aspect of the entity. Here entity is any domain such as products, hotels, movies, etc. And aspect demonstrate the feature of the domain.

2.3 Web Usage Mining

Web mining is the application of data mining techniques to analyse the frequent patterns of World Wide Web. There are 3 types of web mining: Web Usage Mining, Web Content Mining and Web Structure Mining.

Web Usage Mining is a study of user's behaviour on World Wide Web. Web Usage Mining is a concept of extracting web usage pattern and analyse it. There are various application of Web
Usage Mining in real world, like understanding user's behaviour which helps in promoting business, product recommendation, also helps administrator to avoid any illegal activities on web.

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. 5, Issue 6, June 2016

The whole process of Web Usage Mining carried out in 3 phase:

1. Log data pre-processing
2. Pattern discovery
3. Pattern analysis

Figure 2.1 illustrate the Web Usage Mining process.

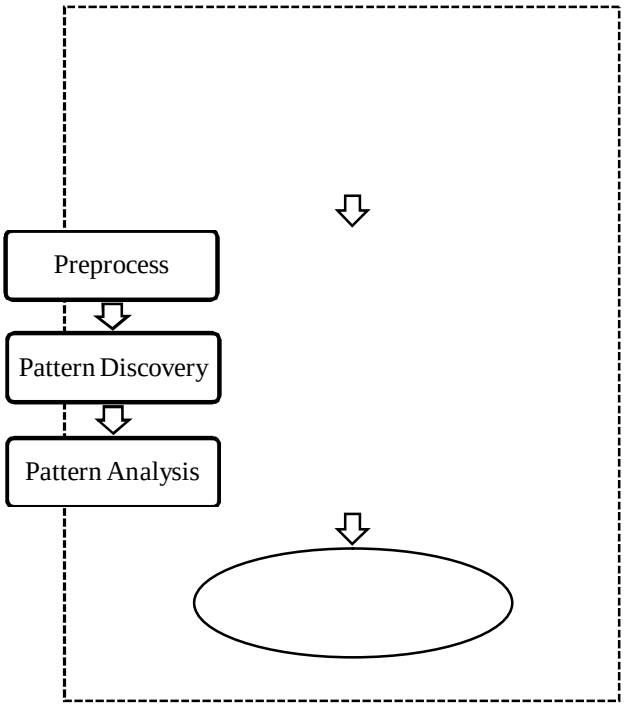

Figure 2.1 Phases of Web Usage Mining

A. Pre-processing

The preliminary step of web usage mining is pre-processing. In this step, web log records are collected and transformed into cleaned log file. In this process all the useless or noisy data are removed and also reduce the size of log file. This cleaned log file is the input to the next phase of web usage mining.

B. Pattern Discovery

As the name, this step is use to discover hidden patterns from cleaned log file. This step make use of different algorithms and techniques of data mining, machine learning, pattern recognition, etc. Clustering, classification, association rule, sequential pattern are well known algorithm of data mining.

C.Pattern Analysis

Final step of web usage mining is pattern analysis. Mined pattern from pattern discovery step is further analysed to extract more relevant and exact patterns to predict user's behaviour.

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. 5, Issue 6, June 2016

III. PROPOSED APPROACH

In this section we have proposed our framework as shown in figure 3.1.The main agenda of our proposed framework is to eliminate the fake reviewers and their reviews and perform product ranking or product classification in terms of their quality. The whole framework is carried out in two phases:

1. Filtering Spam Reviews
2. Product Ranking

Filtering Spam Reviews

Spam reviews are those reviews posted by a fake user or a group of fake user in order to affect the publicity of the product. All the operations performed in this phase is based on the concept of *Web Usage Mining*.
In this phase, web log file is used as an input to the model. A list of activities performed by the user on the internet is automatically recorded and maintained by the server. These list of activities is termed as web log file.

The next step is to clean the log file. All the unwanted entries from the log file is removed and only the relevant entries are kept for future steps. This can be done by observing the extension (.gif, .html, .mp3, .mp4, etc.) of the web files requested by the user which is recorded in the web log file. For our model we kept only those entries having .html extension.

Once we get the clean log file the next step is genuine user identification and their session. In this step, we get the information about the number of user accessing the site. All the users have their individual IP address, which are recorded in web log file, every time when he login to the site. This IP address entries helps to identify the users. Also the same step is responsible for finding the genuine users. To claim a user is genuine or not, here we set a threshold value of 20. If any user visit the site more than the threshold then the model assume that user as genuine. In the log file we will only keep the entries of genuine users.

The next step is to find the session of the genuine user. One can define a user session as a frame of time when a user enter into the site and logged out from the site. A predefine time period (say 30 minutes) is set by the administrator for one session. Any number of time a user can enter and exit from the site within that time period, its session is count as one user session. Once the user re-enter the site after the expire of 30 minutes, then this time a new session is said to be started for the same user and the counter of the session is incremented by 1 and denote two user sessions. In this manner sessions of every users is identified.

The next operation of web usage mining involved in our proposed model is analysis of individual user's choice for different product items.

Finally we have a list of genuine users and their reviews or comments on different products of their interest. These reviews are the input to the second phase of our proposed model.

Product Ranking.

The second phase is to classify or to rank the products in terms of their quality into five categories (very good, good, average, poor, and very poor).The input to this phase is the bag of comments of selected user from genuine user list. All the steps involved in this phase is shown in figure 4.1.The concept of *Opinion Mining* is used to perform the steps.
Initially the dataset are collected from amazon.com. For our proposed model we selected 3 modules: printer, mobile and laptop.

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. 5, Issue 6, June 2016

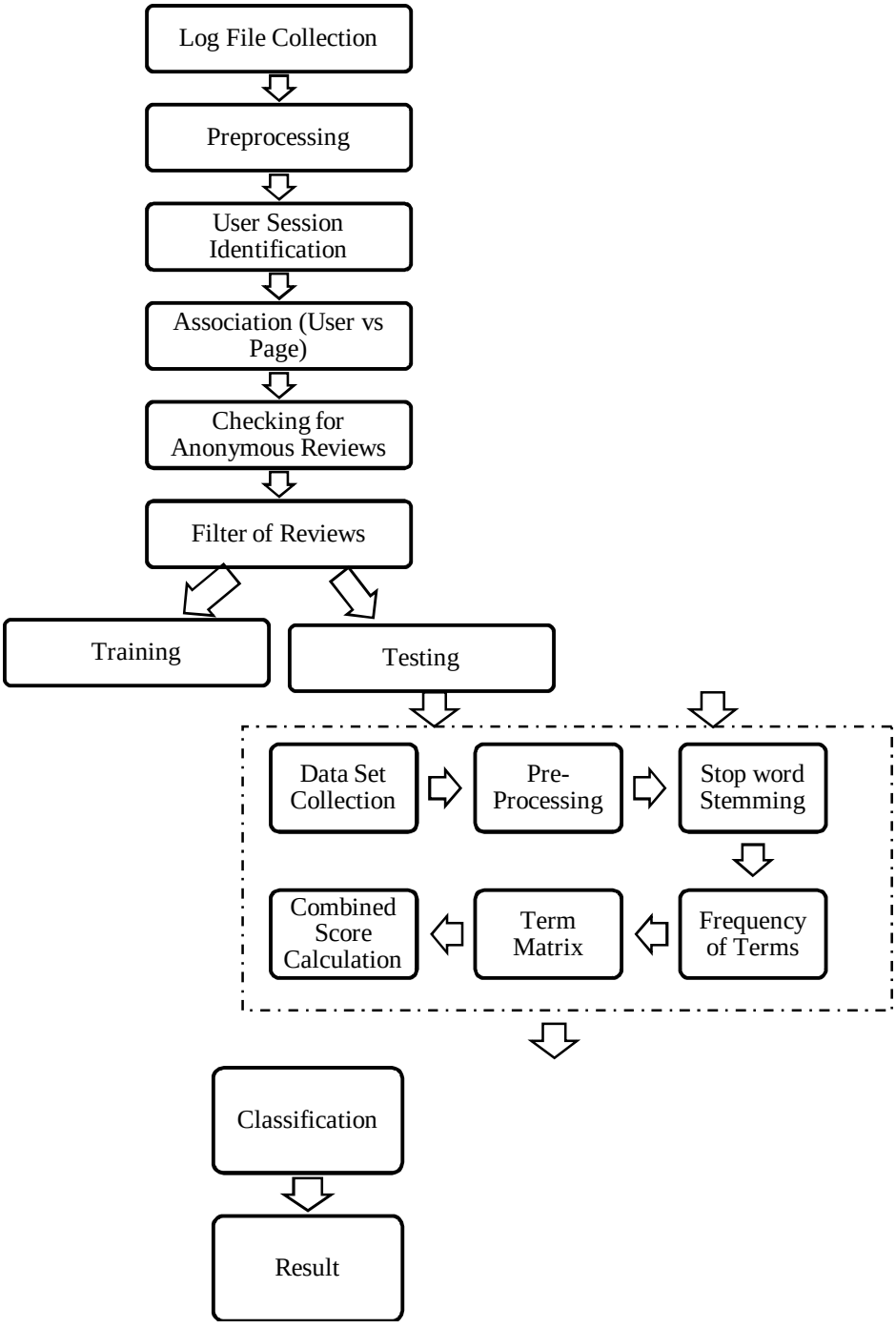

Figure 3.1 Proposed System Architecture

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. 5, Issue 6, June 2016

Once the dataset is created, the next step is pre-processing. The dataset created is not in relevant format. The task of pre-processing is to convert the raw, unstructured, original format data into structured, intermediate data format. There are various algorithms describe in data mining techniques for performing pre-processing step. One well known algorithm for suffix stripping known as *Porter Stemmer Algorithm* is used in our model. Also in our model, we are removing all those supporting words which have no meaning. These words are called stop words and the process can be termed as stop word removal.
The next step is to measure the importance of words or terms within a document. This is done by three factors*: term frequency (tf), inverse document frequency (idf) and term frequency-inverse document frequency (tf-idf).*
Term frequency (tf), counts the number of times each term occur in document.
Inverse Document Frequency (idf) is the measure of how importance the term is .The idf score indicates whether the term is rare or common across all the documents.
The tf-idf is the product of two scores, tf and idf, which indicates how important a term is to a document in a collection.

Next step is building term matrix and combined score calculation. In this step, a matrix of all the terms evaluated and their relative frequency is created. The matrix of terms and the combined score of tf-idf of each terms helps in classification.
Classification is a supervised technique of mapping input data to one of the predefined classes. There are various classification techniques in data mining concept. For our proposed model, we are using Nearest Neighbor Algorithm.
Once classification is done, we have the final result which predict the quality of our product according to the bags of comments of genuine user only.

IV. RESULT AND DISCUSSION

For evaluation, we split our dataset (collected from amazon.com) into training set and testing set. We use 6 products each divided into 3 domains: printer, mobile and laptop. We test our proposed algorithm and compute the accuracy of each domain. In classification problem, confusion matrix is a primary source of accuracy estimation. The table shown below (Figure 4.1, 4.2, 4.3) represent the performance of the proposed classifier on a set of test data.

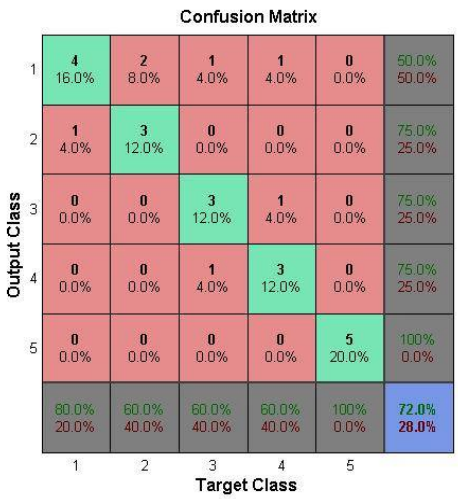

Figure 4.1 Confusion Matrix of Printer

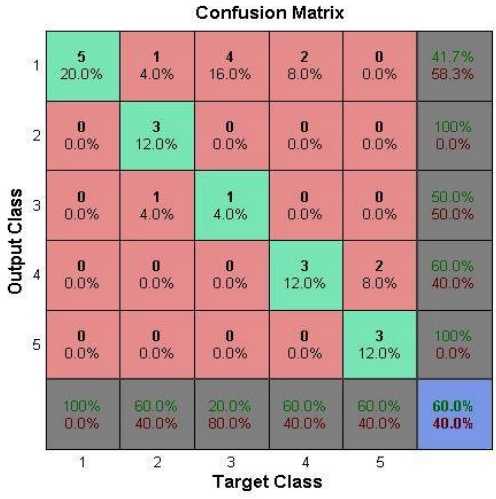

Figure 4.2 Confusion Matrix of Mobile

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. 5, Issue 6, June 2016

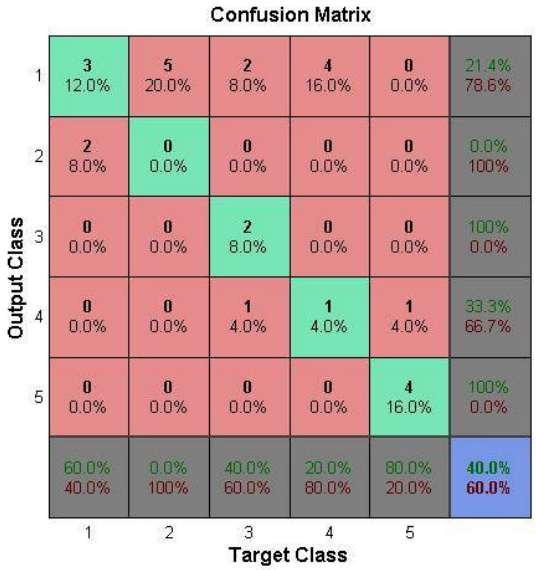

Figure 4.3 Confusion Matrix of Laptop

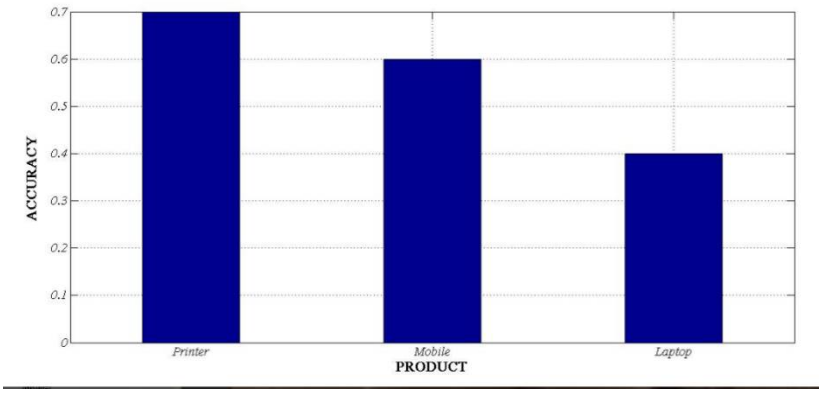

Figure 4.4 Graphical Result of Accuracy of the proposed algorithm

V. CONCLUSION

In this paper, our aim is to improve the ranking system algorithm and also to filter the spam reviews. We use the operations of web usage mining to alleviate the fake reviewer problem and proposed a new method to rank the products according to their quality. To analyse our experiment, we collected the dataset from amazon.com containing the

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. 5, Issue 6, June 2016

consumer reviews of popular products in 3 domains. Experimental results demonstrate the effectiveness of the proposed method.
In future, the main target is to improve the techniques of filtering spam reviews and to adopt real time implementation.

REFERENCES

[1] Alexander Pak and Patrick Paroubek,"Twitter as a corpus for Sentiment Analysis and Opinion Mining".J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Anindya Ghose and Panagiotis G.Ipeirotis,"Estimating the helpfulness and Economic Impact of Product Reviews : Mining text and Reviewer Characteristics",IEEE Transactions on Knowledge And Data Engineering,Vol.23,No.10October 2011.

[3] Zheng-Jun Zha, Jianxing Yu, Jinhui Tang, Meng Wang, and Tat-Seng Chua," Product Aspect Ranking and Its Applications", IEEE transactions on knowledge ans data engineering, Vol. 26, No.5, May 2014.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc. EMNLP, Philadelphia, PA, USA, 2002, pp. 79–86.

[5] W. Jin and H. H. Ho, "A novel lexicalized HMM-based learning framework for web opinion mining," in Proc. 26th Annu. ICML, Montreal, QC, Canada, 2009, pp. 465–472.

[6] Wouter Bancken, Daniele Alfarone and Jesse Davis, "Automatically Detecting and Rating Product Aspects from Textual Customer Reviews", Proceedings of DMNLP, Workshop at ECML/PKDD, Nancy, France, 2014.

[7] Andrea Esuli and Fabrizio Sebastiani ,"SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining."

[8] M.Govindarajan,"SENTIMENT CLASSIFICATION OF MOVIE REVIEWS USING HYBRID METHOD", International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 1, Issue-3, Jan.-2014.

[9] Wouter Bancken, Daniele Alfarone and Jesse Davis, "Automatically Detecting and Rating Product Aspects from Textual Customer Reviews", Proceedings of DMNLP, Workshop at *ECML/PKDD*, Nancy, France, 2014.

[10] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. SIGKDD, Seattle, WA, USA, 2004, pp. 168–177.

[11] Miller G, Beckwith R, Fellbaum C, Gross D, Miller K. WordNet: an on-line lexical database. Oxford Univ. Press; 1990."

[12] G.Vinodhini, RM.Chandrasekaran,"Performance of Hybrid Sentiment Classification Model on Online Product Reviews",International Journal of Computer Science Engineering,IJCSE.

[13] Adam L.Berger, Stephen A.Della Pietra and Vincent J.Deila Pietra 1996. "A maximum Entropy approach to natural Language processing " Computational Linguistic.

[14] Chetan Kaushik and Atul Mishra , "A Scalable , Lexicon Based Technique for Sentiment Analysis" Vol.4 ,No. 5 September 2014, IJFCST.