DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4470

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

FEROM: A PRAGMATIC APPROACH FOR SENTIMENT ANALYSIS AND OPINION MINING

Ms. Dhanashree Raut
Computer Department
MGM's College of Engineering and Technology
Navi Mumbai, India

Prof. Sachin Chavan
Computer Department
MGM's College of Engineering and Technology
Navi Mumbai, India

Abstract: In recent years, online shopping has become popular as it is convenient, reliable and cost effective. The online customer finds it difficult to make purchasing decisions based on the pictures or descriptions provided. The online review often makes it easy for the customer to make decisions for purchasing products as they are a great source to compare products and features. Unfortunately, going through all customer reviews is difficult, especially for popular items as they are in a number of hundreds or thousands. Now-a-days, a large number of availability of rich opinion resources like online review sites and blogs helps customers to understand the opinions of others about the product. We have proposed a system Feature Extraction and Refinement for Opinion Mining (FEROM) which aims to mine customer reviews of a product and extract high detailed product entities on which reviewers express their opinions. The opinions expressed by the customer are reviewed and then they are divided into multiple sentences. The Parts of Speech (POS) tagging is applied on these sentences where each sentence is tagged according to its respective parts of speech. After tagging, their expressions are identified with the help of SentiwordNet dictionary. The words in each sentence are assigned a score and an objective score is calculated with the help of SentiwordNet Dictionary.The sentiment of the customer is identified and opinion orientations for each recognized product entity are classified. These words are then compared with the dictionary of positive and negative which finally segregates the reviews into positive and negative.

Keywords: Opinion Mining, FEROM, POS, SentiWordNet, Feature Extraction, Refinement

I. INTRODUCTION

A number of online customers have increased due to rapid growth of e-commerce and also, due to increase in number of online merchants. To enhance the customer satisfaction, merchants and product manufacturers allow customers to express their review or opinion online on the products or services they use. The customers can express their review on sites like flipkart.com, ebay.in, snapdeal.com and others. These online reviews have become a potential source of information for the customers. Customers have found these reviews helpful in making informed decisions whether to purchase a product. Since customer feedbacks influence other customer's decision, the review documents have become an important source of information for business organizations to take it into account while developing marketing and product development plans

A. The Opinion Mining System

Opinions are central to almost all human beings. They are the key influencers of our behavior. Our opinions to a considerable degree show our perception of reality, and beliefs and the choices we make depend on how others see and evaluate the world. For this reason, we often seek opinions of others while taking any decisions. The opinions play an important role for both customers as well as manufacturers. Opinions and concept related to it such as sentiments, attitudes and emotions have become subject for study of sentiment analysis and opinion mining [1]. The rapid growth in this field coincides with those of the social media such as reviews [2] [3], blogs, discussion forums, twitter, social networking etc. because for the first time in history we have a huge volume of opinionated data recorded.

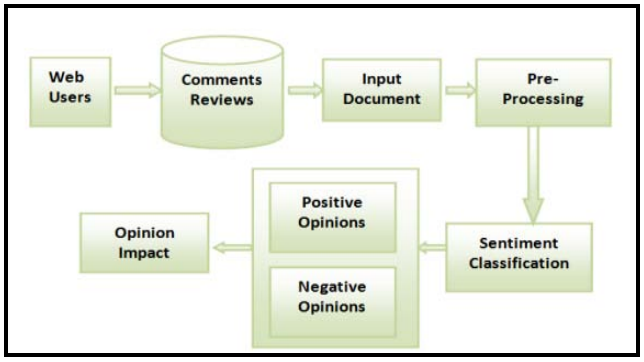

Figure 1. Opinion Mining System

The opinion mining system in Figure 1 gathers the review from online sites and cleans the reviews. The review crawler cleans the review using various methods and the data processor processes the sentence using the algorithm. The algorithm removes the noisy reviews that are unwanted. The parser splits the sentence according to parts of speech and categorizes it into verbs, nouns, adjectives etc. The features and opinions expressed by the reviewer are identified and the sentiments it conveys are recognized. The feature opinion pair is obtained and categorized which tells us about opinion expressed by the reviewer. The existing approaches proposed by various authors removes the noisy reviews from the sentences that are written in UK English but does not focus on US English which most of the users use while writing reviews. Also, the words that have the same meaning are considered as different terms and this result in incorrect opinion analysis. Our proposed system Feature Extraction and Refinement for Opinion Mining (FEROM)

overcomes these issues and also works on reviews written in US English to remove unwanted terms and slangs. FEROM deals with the informal opinions of the customer. The system also merges the terms with same or similar meaning and result in correct opinion analysis. Also, the reviews are segregated into positive and negative.

B. Previous Work

A substantial work has been done on sentiment analysis for reviews and to determine the subjectivity of sentence. Another area related to this domain feature based sentiment analysis which determines the opinion on the basis of related attribute. Most of the work concentrates on identifying sentiments of a reviewer. The study of various authors describes various schemes and techniques used for opinion mining and sentiment analysis.

Erik Cambria et al [1] focused on opinion mining and sentiment analysis technique based on NLP and data mining. The authors discussed the history, current use and future of opinion mining and sentiment analysis. They also focused on common sentiment analysis and evolution of opinion mining. Their work helps in identifying pro and con expression of online review and increases the credibility of product evaluation. The focus of opinion mining and sentiment analysis is mainly on polarity detection and emotion recognition. Several tools exists which rely on parts of text and thus, fail to capture the opinion or sentiment of the reviewer. The evolution of opinion mining and sentiment analysis rely on vector extraction which is used for representing the most important and salient features of the text. This vector can be used for classification of the most relevant features. The authors used term frequency and presence as common features. Presence is a binary-valued feature vector whose entries indicate whether the term occurs or doesn't. The term frequency is a vector which indicates the occurrence of term in a sentence. This system obtained better performance using presence than term frequency. A topic may be more highlighted using frequency of occurring of a term but the overall sentiment of a sentence cannot be classified using frequently occurring terms. These approaches are bound by a particular language domain and thus applying these techniques to other language domains is a problem. **Kunpeng Zhang et al. [2]** presented a product ranking model that applies weights to review factors in order to calculate product ranking score and produce results closely related to sales ranking as reported by the retailer. **Ramanathan Narayanan et al. [3]** have proposed a feature based product ranking technique which first identifies product feature within a product category and analyzes their frequency and relative usage. For each feature, subjective and comparative sentences are identified and sentiment orientations are assigned to these sentences. The relative quality of the products is determined using the weighted and directed graph constructed with the help of information obtained from the customer reviews. **G. Di Fabbrizio et al [4]** have proposed a novel method STARLET for extracting reviews of a product in a more summarized manner based on rating distribution and leads to more readable reviews. In this scheme, author considers a particular domain for example restaurant and aspects for example atmosphere, food, service to score each sentence in input documents. For each aspect, STARLET uses a maximum entropy rating model to calculate how much a particular sentence has contributed to that particular aspect. The predicted aspect ratings are used to calculate a score for each sentence and also to derive a summary score. The version of STARLET used by the authors did not have a mechanism to avoid similar sentences so there was a slight problem of redundancy. **Hideki Asoh et al in [5]** proposed a context-aware personalized movie recommender system for mobile phone users. The authors have used a Bayesian network for recommendation of movies on mobile phone based on the users' profiles, histories, situations, and movie contents. The approximate ranking method was used to reduce the computational cost for probabilistic inference of Bayesian network. The system recommends a list of movies to mobile users' based on their profile, history, situation and the person they accompany. This kind of system can be applied to areas such as e-commerce, shopping malls**. Yin-Fu Huang and Heng Lin in [6]** have proposed a system which calculated the product score for ranking of web products using opinion mining. The ultimate goal of the system was to find the favorable or interesting product for an individual from amongst a huge amount of products. This system while calculating the product score considered three issues the product reviews, product popularity and the product release month. The opinion mining technique was used to identify the sentence polarity and then the product score was calculated of all matched products. This system can be used to find the distribution of top-ten products according to release year and their specified features. These results also revealed that new products are not always favorable than the old products. **Weishu Hu et al. [7]** have proposed an extraction approach which only mines the features of the product in the opinion sentences. The SentiWordNet based algorithm was proposed by the authors to find the opinion sentences. Their approach involves identifying the opinion sentences in each review using SentiWordNet as positive or negative. These sentences are then used for mining product features that the customers have commented on. After this the incorrect features are then removed by pruning. This scheme uses data mining and natural language processing both for feature extraction. This gives higher precision and recall values as compared to previous schemes. **Lizhen Liu et al. [8]** have proposed a novel method to deal with the problems of feature-level opinion mining. The proposed method considered the explicit features and implicit features for opinion mining and the opinion words are divided into vague words and clear words. After this clustering of features is carried out this depends on the corresponding opinion words, structure of the feature and similarity of the feature. The context information was used which strengthen the procedure of clustering and helped in filtering the noise. This scheme has shortcoming such as small corpus cans cannot perform well and the structure of the dictionary increases the cost of the method. **Liu Gongshen et al. [9]** have proposed a new model for predicting semantic orientation of a review. This model categorized positive review from negative ones by integrating grammatical knowledge and taking into account topic correlations. The reviews are parsed using Stanford parser and the adjectives and the features are extracted using their grammatical relationship. Then based on WordNet polarity of adjectives is predicted. The similarity of feature and topic is calculated. The value of every noun is calculated based on polarity and similarity obtained. To

calculate the semantic orientation a vector is defined with a weighted score and the polarity labels are produced which categorizes reviews into positive and negative.

Heng-Liyang and Qing-Fenglin [10] have proposed a scheme for multi-scenario problems, which collects review from movie review sites and infers the authors' sentiment using evolutionary computational strategies. It helps to compute characteristic value and opinion intention value from specific scenario. The authors collected reviews about a movie from a Taiwan electronic bulletin board. These reviews consist of some sentences with unknown words for example urban words. The authors removed the sentences containing the unknown words and fed the remaining sentences to a free Chinese word segmentation system to get segmented into words for further processing. The average judgment score for every sentence was calculated by averaging the judgment of participants. After this ConceptNet, a general common sense ontology dictionary based on daily life English was used to find out the characteristic value of each sentence. Then the opinion intention value for a sentence B at a specific scenario A was calculated by summing up the product values obtained from taking the cross-product of the weight table of scenario A and the transposed characteristic values of sentence B. Thus, evolutionary computing strategies were applied by authors to optimize the weights for multi-scenarios. This approach has better correct rates compared to previous approach and it not only helps in solving multi-scenario problems but also helps to understand which emotional elements are important and which can be discarded. The scheme proposed in **[11]** integrates a semi-supervised learning algorithm that is used for text polarity classification with a RapidMiner, a platform for predictive anylsis. The authors examine an approach used for building a graph based semi-supervised sentiment polarity classifier. The text polarity classification is carried out based on the knowledge base provided for opinion mining.In this scheme for text polarity classification, a dictionary of their own is created by authors which is used for tagging the words from the example set documents. The preprocessing is done by Rapid Miner Text Input operator chain which is given as input to the SentimentDictionaryLoader.The goal of SentimentDictionaryLoader is to load the set of desired sentiment word sets. The semi-supervised learner has two inputs the example set and the word polarity list and after applying the conjugate gradient method a classifier is provided based on sentiment dictionary which provides both labeled and unlabeled text documents as the output.The unlabeled data cannot be used for validation scheme because it is auxiliary and its labels are unknown. This may affect the validation results,so the unlabeled data is fed to a semi-supervised learner and the cross-validation is applied.This scheme has taken into account both the memory usage and the processor.This scheme was extension for frequently used open source data mining tool and was consider valuable as semi-supervised learners were provided with this tool.

II. PROPOSED SCHEME

The purpose of review analysis is to extract, organize, and classify the information contained in the required documents. The required document undergoes cleaning for unwanted stop-words and words not listed in the dictionary. It is then classified as formal or informal depending on the use of words, which then undergoes the application of POS tags. Thus the feature and its accompanying opinion is identified and extracted. FEROM generates feature-opinion pairs from the review documents which segregates the reviews into positive and negative.

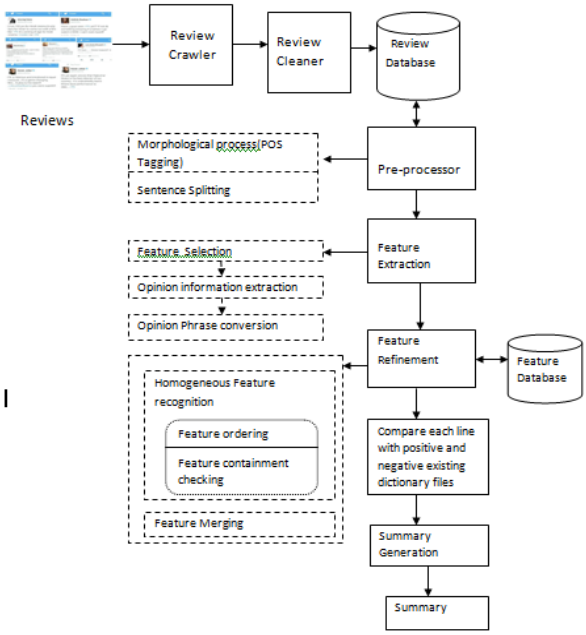

Figure 2. Feature Extraction and Refinement for Opinion Minng (FEROM)

The propose method is an enhanced method called, feature extraction and refinement for opinion mining (FEROM). The architecture of the proposed opinion mining system FEROM as shown in Figure 2 consists of various functional components as given below:

- review documents crawler,
- document pre-processor,
- document parser,
- feature and opinion learner

1) Review Documents Crawler and Review Document Cleaner

For a target review site, the crawler retrieves review documents or review data from online stores and stores them locally after filtering markup language tags. The review cleaner removes unnecessary content such as HTML tags and then stores the review data to the review database.

2) Document Pre-processor

The pre-processor conducts morphological analysis of the review data including POS tagging, splits a compound sentence into multiple sentences, and performs stop word removal and stemming.

a) Morphological Analysis

The filtered review documents are divided into manageable record-size chunks whose boundaries are decided heuristically based on the presence of special

characters. It has been found that granularity of words, word stems, and word synonyms may cause problem while extracting real features and opinion. We have applied rigorous reprocessing on review documents to filter out noisy reviews that are introduced either without any purpose or to increase/decrease the popularity of the product. The sentence is scanned according to Parts of Speech and divided according to that.

b) Sentence Splitting

The scanned compound sentences are divided into multiple simple sentences. This makes the morphological analysis easy.

3) Feature Extraction

The feature extractor extracts product features from pre-processed review data. Feature extraction proceeds in three phases:

b) Feature Selection

It selects a candidate feature in a sentence by looking for a noun phrase.

c) Opinion Information Extraction

It finds an opinion phrase that is associated with the candidate feature, and

d) Opinion Phrase Conversion

It replaces an opinion phrase expressed using a negative term with its antonym.

4) Feature Refinement

This feature refinement reduces the number of features by merging candidate features with the same or similar meanings, defined as homogeneous features the feature refiner is divided into two parts

4.1) Homogeneous Feature Recognition

a) Feature Ordering

The feature refiner recognizes homogenous features by exploiting the feature ordering process that synchronizes the word orders of the features to detect synonymous feature candidates.

b) Feature Containment Checking

The feature containment checking is a process that examines the subset superset relationship between the features to check for similarity between them.

4.2) Feature Merging

Finally, the feature merging process merges homogeneous features into a representative feature and also prunes the feature candidates that have significantly low frequencies and very small amounts of related opinion information.

5) Comparison with the Dictionary

The words and sentences from the feature merging module will be compared with the dictionary of positive and negative words and then the output will be given.

III. IMPLEMENTATION OF FEROM

The system contains five modules which are Review document crawler, Tagging, Feature Extraction, Orientation Identification and Sentiment Analysis and Summary Generation as shown in Figure 3. The Review document crawler module gathers the review [4] [5] and the Review Classification unit under this module classifies the review. The Tagging module splits the sentences into multiple sentences and performs Parts of Speech (POS) tagging on these sentences. The Feature Extraction module builds a feature list and then perform grouping on this list which identifies the similar features and removes redundancy. The Orientation Identification module identifies orientation of each word into positive, negative or neutral based on the features provided by the user and the grouping done in previous module. Similarly, Orientation Identification is performed for Opinion Sentences in which whole sentence is identified into positive, negative or neutral. The Sentiment Analysis module analyses whether the given review is positive or negative with a score given to each sentence based on its orientation as positive, negative or neutral. The Summary generation module generates the summary which gives us the result that the reviews for the given product are positive, negative or neutral.

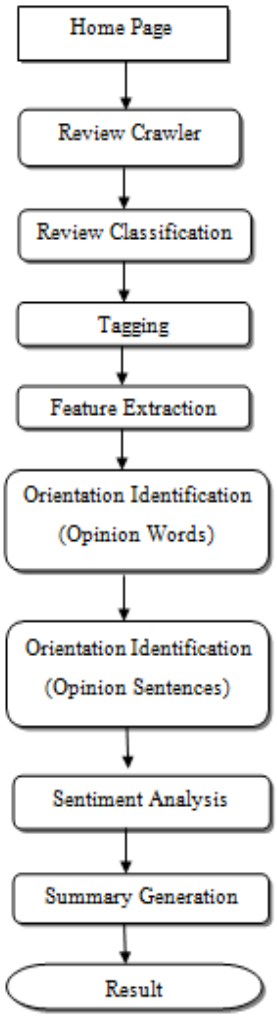

Figure 3. Flow of FEROM

A. Review Crawler

The Review Crawler fetches the webpage after getting the URL and retrieves the reviews from a target review website. Also, it extracts all the words from the page and saves them in database and also saves the order of the words

so that it helps user in searching words and not just keywords.

B. Review Classification or Pre-processing

The review documents are then divided into record sized chunks after removal of all the mark-up language tags. The records size chunks have boundaries that are decided by the special character. The Review classification component uses pre-processing to filter noisy reviews that are introduced to increase/decrease the popularity of product or introduced without any reason. It also removes the special characters that are used while writing the reviews.

C. Tagging

The Tagging component uses Parts of Speech(POS) tagging which means assigning grammatical tags with each word of the pre-processed sentence .Assigning grammatical tags means assigning parts of speech to each word of the sentence and then the tagged output is produced. In FEROM Stanford NLP parser is used as POS tagger in which words are separated depending on their parts of speech.

D. Feature Extraction

The feature extraction component extracts features [6] [7] from the pre-processed reviews. A feature is selected based on the noun phrase in the sentence related to it, and then an opinion phrase is selected that expresses the opinion of the candidate about the product. If any negative [12] phrase is encountered then the phrase is replaced by its antonym. A list of features is generated using this approach for all the pre-processed reviews [13].

E. Orientation Identification(Opinion Words and Sentences)

In orientation identification component the features extracted are given as input and the orientation of these words is identified by the system. The words are compared with the SentiwordNet dictionary and a score for each word is calculated. These words are compared with the negative and positive list of words that is stored.

F. Sentiment Analysis

After the identification of the sentences into positive, negative or neutral based on the words and the sentences that are identified as positive or negative a score is generated using the Sentiment analysis [14] [15] algorithm. Then the sentence is segregated as positive, negative or neutral based on these scores. Thus, we conclude that the review given by the user is positive, negative or neutral.

G. Summary Generation

At the end a summary is generated based on all the positive and negative sentences and their scores are calculated. These scores shows that how many reviews were positive and how many reviews had negative opinions and an overall result is generated after evaluation of all the reviews that there were more positive reviews or negative reviews or neutral reviews.

IV. RESULT AND DISCUSSION

The Result analysis is measured based on some parameter. The parameter considered here is Accuracy which is used to compare FEROM and the existing system. It compares how accurate the informal reviews are segregated from the formal reviews.

A. Accuracy

The FEROM is used to retrieve reviews from the website and segregate those reviews into Positive and negative which helps user in making decision opinion about the product and helps in decision making. The existing system did not work on informal reviews which have grammatical errors and spelling mistakes. It also included similar meaning words as noise which produced improper results which affects the accuracy of the system during analysis.

FEROM deals with these informal reviews and removes the grammatical errors and spelling mistakes. It also removes slang words that are mentioned in the review. Also, it groups the similar meaning words into one group which gives more accuracy during analysis than the existing system. Figure 4 shows the accuracy of FEROM is more than Hyperlink-Induced Topic Search (HITS).

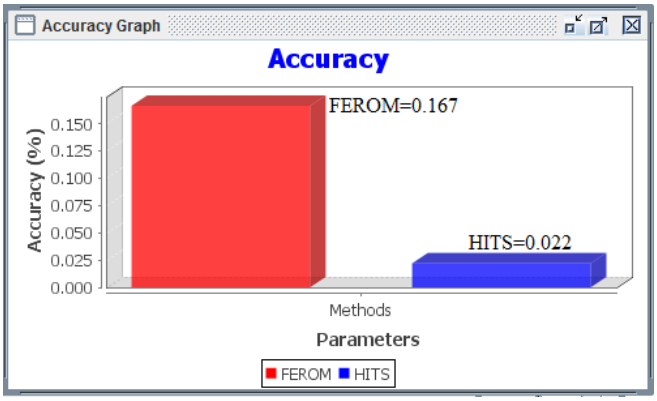

Figure 4. Accuracy Measurement Chart

B. Comparison of Reviews

The comparison graph shows the comparison of positive, negative and neutral reviews. The chart shows which type of reviews product has and thus, gives us the feedback of the user. If there are more positive reviews than negative and neutral, then the positive bar will be highest thus the reviews are positive and positive feedback. If the negative reviews are more than the negative bar will be highest which results in negative reviews and negative feedback. If neutral reviews are more as compared to positive and negative reviews then the neutral bar will be high resulting in neutral feedback as shown in Figure 5.

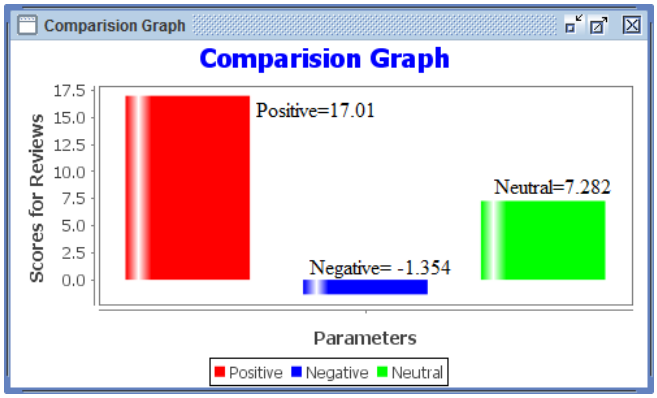

Figure 5. Comparison of Reviews

V. CONCLUSION

Opinion Mining refers to the extraction of opinions and feature about a product from the reviews. The existing system does not deal with the informal reviews and it also includes the similar meaning words which gives improper analysis of opinion mining. The relationship between the term and the related opinion is not studied, for example, only term frequency is considered and not the relation between the term and opinion. A positive review does not mean that the user has positive review on all aspects of the product. Similarly, for a negative review about the product does not mean that the user has negative review on all aspects related to the product. Our proposed system FEROM extracts features from the review and considers the syntactic and semantic similarities and segregates the reviews into positive and negative. FEROM deals with the informal reviews which contains use of improper English, grammatical errors and spelling mistakes. It removes slangs or short words, and also removes the similar meaning words that act as noise while analyzing the sentiment of the given review. FEROM analyses the relation between the noun phrase and the opinion phrase and then segregates the reviews into positive and negative. The experimental results show that accuracy of FEROM is better the existing approach Hyperlink-Induced Topic Search (HITS).

In future, opinion mining systems need to become broader and acquire deep knowledge bases. We will try to improve the efficiency by combining the reasoning methods that are inspired by human thoughts and psychology. This will help in better understanding of natural language opinions and also help to bridge gap between unstructured and structured data processing. Also, we will make our system to work for analysis of multi-lingual reviews. The system will help in segregating reviews on a site which are in different languages and a combined result can be given on the basis of all the reviews. Finally, we will try to make our system more intelligent by blending emotional theories with practical goals of analyzing the natural language reviews so that the system can make analogies, learn new affective knowledge, and can detect, perceive and feel emotions.

VI. ACKNOWLEDGMENT

We are thankful to MGMCET who gave us the opportunity to complete the research work. We would like to express our gratitude to our Professors, who guided us with their professional experiences and knowledge. Also, we would like to thanks our colleagues, friends and family for providing us useful comments, suggestions and continuous encouragement.

VII. REFERENCES

[1] Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," IEEE Intelligent system, Vol. 28, No. 2, March-April 2013, pp. 15-21.

[2] Kunpeng Zhang, Yu Cheng,Wei-keng Liao, Alok Choudhary, "Mining Millions of Reviews: A Technique to Rank Products Based on Importance of Reviews," in Proceedings of the 13th International Conference on Electronic Commerce, August 2011.

[3] Ramanathan Narayanan ,Kunpeng Zhang, Alok Choudhary, "Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking" in Proceedings of the 3rd conference on Online Social Networks, June 2010.

[4] G. Di Fabbrizio, A. Aker, and R. Gaizauskas, "STARLET: Multi-Document Summarization of Service and Product Reviews with Balanced Rating Distributions," 11th IEEE International Conference on Data Mining Workshops, 11-11 Dec. 2011, pp. 67–74.

[5] Hideki Asoh,Chihiro Ono,Yoichi Motomura, "A Movie Recommendation Method Considering Both User's Personality and Situation," Workshop on Recommender Systems, Riva del Garda, Italy, August 28-29, 2006.

[6] Yin-Fu Huang, Heng Lin, "Web Product Ranking Using Opinion Mining," IEEE Symposium on Computational Intelligence and Data Mining, 16-19 April. 2013, pp. 184-190.

[7] Weishu Hu, Zhiguo Gong, Jingzhi Guo, "Mining Product Features from Online Reviews," 7th IEEE International Conference on E-Business Engineering, 10-12 Nov. 2010, pp. 24-29.

[8] Lizhen Liu, Zhixin Lv, Hanshi Wang, "Opinion Mining Based on Feature-Level," in 5th IEEE International Congress on Image and Signal Processing, 16-18 Oct. 2012, pp. 1596-1600.

[9] Liu Gongshen, Lai Huoyao,Luo Jun, Lin Jiuchuan, "Predicting the Semantic Orientation of Movie Reviews," Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 10-12 Aug. 2010, pp.2483-2488.

[10] Heng-Liyang , Qing-Fenglin , "Sentiment Analysis In Multi-scenarios: Using Evolution Strategies For Optimization," Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 2013,pp. 1230-1233.

[11] Balla-Müller Nóra, Camelia Lemnaru, Rodica Potolea"Semi-Supervised Learning with Lexical Knowledge for Opinion Mining," IEEE 6th International Conference on Intelligent Computer Communication and Processing, 26-28 August 2010.

[12] D. Olsher, "Full Spectrum Opinion Mining: Integrating Domain, Syntactic and Lexical Knowledge," 12th IEEE International Conference on Data Mining Workshops, 10-10 Dec. 2012, pp. 693–700.

[13] Gamgarn Somprasertsri, Pattarachai Lalitrojwong, "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization," Journal of Universal Computer Science, Vol. 16, No. 6, 2010, pp. 938-955.

[14] Haseena Rahmath P, "Opinion Mining and Sentiment Analysis - Challenges and Applications**,"** International Journal of Application or Innovation in Engineering & Management, May 2014, Volume 3, Issue 5,pp. 401-403.

[15] G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, June 2012, Vol.2, Issue 6, pp. 282-292.