International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 6 3894 - 3900

Exploiting and Ranking Dominating Product Features through Communal Sentiments

Mr. S. P. Ghode
Department Of Computer Engineering
DGOI, FOE, Daund
Savitribai Phule Pune University, Pune
Daund, Pune, India
shyamghode@gmail.com

Prof. S. S. Bere
Department Of Computer Engineering
DGOI, FOE, Daund
Savitribai Phule Pune University, Pune
Daund, Pune, India
sachinbere@gmail.com

Abstract- The rapidly expanding e-commerce has facilitated consumers to purchase products online. Various brands and millions of products have been offered online. Varieties of customers' reviews are available now days in internet. These reviews are important for the consumers as well as the merchants. Most of the reviews are disorganized so it generates difficulty for usefulness of information. In this paper we are proposing a product feature ranking framework, which will identify important features of products from online customer opinions, and aim to improve the usability of the different reviews. The important product features are recognized using two observations 1) the important features are mostly commented on by a large number of users 2) users reviews on the important features are greatly influence on the overall reviews on the product. We first identify product features by shallow dependency parser and determine customer's reviews on these features via a sentiment classifier. Then we adopt develop a probabilistic feature ranking algorithm to conclude the importance of features by considering frequency and the influence of the influence of the users reviews given to each feature over their overall reviews.

Keywords- Sentiment *Classification, Naive Bayes, Feature Mining, Feature Clustering, Feature Ranking*

*****

I. INTRODUCTION

In recent years we have observed a rapid growth in e-commerce and online product selling websites. Recent studies on retail spending have estimated $37.5 billion in Q2 2011 U.S. [5]. Different kinds of products by various merchants have been offered online for online consumers. For example, Bing has placed more than five million products to be sold. Amazon.com has a total of more than 36 million products online [2]. Shopper.com has up to five million different products to sell from over 3,000 vendors. Most retail websites allows consumers to write feedback reviews or opinions on various products or on aspects of the products to express their sentiments towards the product. A product feature refers to a component or a specification attribute of a certain product. A sample consumer review may include a sentence like *"The battery life of Nokia N95 is amazing."* indicates positive sentiment on the feature "*battery life*" of product Nokia N95. Excluding the e-commerce and online product selling websites, many other places like forum websites also provide a place holder for consumers to post their reviews on products. For example, CNet.com has more than seven million product reviews; whereas Pricegrabber.com owns millions of reviews on more than 32 million products in 20 distinct categories over 11,000 sellers. Such consumer reviews contains rich and valuable information and have become an important resource for both consumers and firms [3], [9]. Consumers may seek product information from those reviews before purchasing a product, while many firms and vendors use those reviews as an important feedback in their product manufacturing, evaluation, marketing, and consumer relationship management.

A product may have a numerous features or specifications. Such as iPhone 3GS has up to three hundred aspects such as "usability", "exterior design", "Bluetooth", "3G network". These features have different levels of importance over consumers' purchase decision and also over vendors manufacturing strategies. For example, some aspects of iPhone 3GS, e.g. *"processing speed", "RAM"* and *"battery"* are taken into consideration by most consumers, and has more importance than the others such as *"USB"* and *"Button"*. For a camera product, the features such as *"lenses"* and *"picture quality"* would highly influence consumer purchase decision and they are more important than the *"A/V cable"* and *"wrist strap"*.

Figure 1: Features of a typical Smart Phone

Hence, exploiting dominating product features will improve the usage of reviews and is beneficial to both consumers and manufacturing firms. Those product reviews

then further can be analyzed by the consumers' to make a wise decision about purchase while the vendors may have enough knowledge about the product quality issues and to make the required changes to the product. However, it is eventually an overhead and critical task for consumers' to identify the important features by reading the reviews manually and analyzing the public sentiments over these features.

Taking considerations from the above observations, we propose a framework to automatically identify the dominating product features from online consumer reviews and ranking them with their public sentiments. We assume that the dominating product features has following characteristics along them: (1) they occurs more frequently in consumer reviews; and (2) consumers' sentiments over these feature highly affect the purchase decision. A basic approach to exploit the influence of consumers' sentiments on specific feature over their overall ratings on the product is to count the cases where their sentiments are consistent, and then ranks the feature according to the number of the consistent cases.

II. RELATED WORK

Analysis of public sentiments is an extension of data mining and information retrieval which involves processing natural language and extracting information for the purpose of obtaining users' positive, negative and neutral emotions by analyzing enormous amount of data [9]. If we deal with terms for sentiments detection, considering the pitch of the voice, tone of voice, attitude of the speaker this are the features which are involved in it.

As the internet is expanding the text based analysis of sentiments is attracting the researchers' attention. Sentiment analysis classifies the text corpus into positive, negative and neutral categories [10]. Opinion Mining and Opinion Analysis refers to the problems related to products feedbacks, political posts, news groups, reviews sites etc. [11]. There are various methods for summarizing of customer reviews like Text Classification and Text summarization [12]. In earlier days before purchasing any product customers used to ask the reviews to his family and friends to take purchasing decision while retailers needs to decide about their products to improve quality of services, they conduct surveys to the focused groups [13].

Sentiment mining can be categorized in different levels like Review level, Sentence level and Phrase Level.

Depending on which type of data is to be analyzed the type of sentiment analysis level may be selected. Various approaches are considered in literature for Sentiment Analysis and Classifications and are divided into two domains as Machine Learning Approach and Lexicon Based Approach. Further they again can be divided into subcategories they are listed below.

1) *Machine Learning*

Machine Learning Approach is mainly divided in to three categories supervised, unsupervised and semi supervised. Each category again sub divided in to different Machine Learning Algorithms.

a) *Supervised Learning*

Supervised Learning or Classification is used for predicting the result from the given set of values on the basis of defined set of attributes and given predictive attributes. Training data contain the attributes without having prediction attributes [14]. At the testing phase accuracy is check by how accurate the machine is predicting the values. Supervised learning again sub-categorized like Support Vector Machine, Naive Bayes, Maximum Entropy, Decision Tree etc. [15]

The Following table consist the models in supervised learning.

Name of model	Used Learning algorithms
Experiments with SVM to classify opinions in different domains	Support Vector Machines
Feature Based Opinion Mining of Online Free Format Customer Reviews Using Frequency Distribution and Bayesian Statistics	Naïve Bayes
Sentiment Identification Using Maximum Entropy Analysis of Movie Review	Maximum Entropy method
Document-level sentiment classification: An empirical comparison between SVM and ANN	SVM classifier ANN classifier
Which Side are You on? Identifying Perspectives at the Document and Sentence Levels	SVM Naïve Bayes
Automatic Sentiment Analysis of Twitter Messages	Naïve Bayes

Table1: Supervised Learning approaches

b) *Unsupervised Learning*

Unsupervised Learning does not require training instances. It uses different clustering algorithms like Partition based clustering and Hierarchical clustering to classify data into classes. Neural Network can be used for defining threshold values of the words and then classify

them according to the defined values. Semantic Orientation and Point wise mutual information is also used for the unsupervised classification in sentiment analysis [16].

Name of model	Used Learning algorithms
A Novel Product Features Categorize Method Based On Twice-Clustering	K-Means COP K-Means
A Framework To Answer Questions Of Opinion Type	Bayes classifier k-means clustering
An Unsupervised Method For Joint Information Extraction And Feature Mining Across Different Web Sites	undirected graphical model
Enriching Sentiment Polarity Scores Through Semi-Supervised Fuzzy Clustering	Fuzzy-C clustering

Table 2: Unsupervised Learning approaches

c) *Semi Supervised Learning*

Semi supervised approach combines supervised and lexicon based approaches. By using this combination the system performance get improved for classification, because it will give the word stability and readability from a lexicon based approach and high accuracy from the supervised approach [17].

2) Lexicon Based Approach

Sentiment words are always classified in positive and negative categories. There also present opinion phrases and idioms which collectively called as opinion lexicon [17]. There are three approaches, first is manual which is time consuming other two which are automated, they are dictionary based and corpus based. In dictionary based approaches a small set of sentiment words are collected manually, this will grow by searching their thesaurus for their synonyms and antonyms [4], [18]. The newly found words are then added to the list and again the iteration starts. It will stop when no new word found. The drawback of this approach is the inability to find sentiment words with domain and context specific. Corpus based approach finds the sentiment words with context specific manner. It uses two sub categories as follows:

1. Statistical approach: It finds co-occurrence pattern or by seeding sentiment words.
2. Semantic approach: It gives an opinion value which directly relies on different principles for computing the similarity between the words.

III. PROPOSED SYSTEM

We first identify product features in the reviews by Part Of Speech Tagging. We adopt Stanford Parser[1] as a **POS (Part Of Speech) Tagger** [19] and then analyze consumer sentiments on these features via a sentiment classifier. We then adopted a probabilistic aspect ranking algorithm, which effectively exploits the aspect frequency as well as the influence of consumers' opinions given to each aspect over their overall opinions on the product in a unified probabilistic model. In particular, we assume the overall sentiment in a review is generated based on a weighted aggregation of the opinions on specific aspects, where the weights essentially measure the degree of importance of these features. A probabilistic ranking algorithm is developed to infer the importance weights by incorporating aspect frequency and the associations between the overall opinion and the opinions on specific feature [1].

Product Feature Ranking was first introduced in our previous work [20]. Taking in consideration the preliminary conference version [20], here we propose the following improvements: (a) to elaborate more discussions and analysis on unique feature identification problem; (b) to perform extensive evaluations on more products in more diverse domains; and (c) to demonstrate the potential of aspect ranking in more real-world applications.

A. Preliminaries

Part-Of-Speech Tagging: The **Part-Of-Speech Tagging (**POST) refers to naming each word within a given sentence with its appropriate grammatical tag. Stanford NLP parser of Dependency parsers' are used to POST a sentence.

Sentiment Analysis: Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Sentiment Classification: Sentiment Classification is a process of polarizing a given textual data to positive, negative or neutral one. The classification can be done using different text based classifiers.

Opinion Mining: The opinion mining refers to collect different types of opinions of people over the same issue from web. Mostly the reviews are the best way to express users' opinion on web.

Feature Extraction: A feature is referred as a property of something over which people tends to discuss. The features can be extracted by using Natural Language (NLP) Techniques such as Part-Of-Speech (POS) Tagging. Generally the Nouns and Noun Phrases are usually returned as feature. Sometimes feature also replaced with name aspect.

Clustering: The process of grouping elements with relevant properties refers to Clustering. Different types of text clustering are investigated in literature such as K-Means, KNN, K-Medoid, Single Link and Average Link. The distance between the groups of feature vectors decides the clustering criteria.

Probabilistic Ranking: The Ranking refers to arranging the results set as to improve the usability at the user end. The Ranking needs to be considering the ranking criteria for example R-Score (Relevance Score). Probabilistic Ranking takes into account the overall influence of some properties and the probability that the property may be present in the result set.

B. The Framework

1. *Loading Pros and Cons:* First of all we have to load the 'Pros' and 'Cons' statements. The statement usually provided with XML or TEXT version. We have to design a XML or a TEXT parser to read the statement serially. These statements should be loaded into arrays of string data type.

2. *The Naive Bayes Classifier:* As we know the 'Pros' and 'Cons' statement provides the sentiment as positive and negative over an issue respectively. Here we propose to ***train*** a Classifier i.e. Naïve Bayes as a ***Sentiment*** Classifier. The string arrays are loaded as training instances. Naïve Bayes creates a trained model. This classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

Training Time: O ($|D|L_d + |C||V|$))

Where, L_d is the average length of a document in D and V is set of words in D.

Assumes V and all D_i , n_i, and n_{ij} pre-computed in O ($|D|L_d$) time during one pass through all of the data.

Generally just O ($|D|\ L_d$) since usually $|C||V| < |D|\ L_d$

Test Time: O ($|C|\ L_t$)

Where, L_t is the average length of a test document.

3. *Part-Of-Speech Tagging (POST):* A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), ***such*** as ***noun***, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. We use open source Stanford NLP parser for POST. The parser is instantiated with English Model.

4. *Feature Identification:* For the Pros and Cons opinions, we will identify the features by extracting the frequent noun terms in the reviews. For identifying features in the free text ***reviews***, a solution is to employ an existing feature identification approach existing approach ***that*** first identifies the 'Nouns' and 'Noun phrases' in the reviews. The occurrence frequencies of the nouns and noun phrases are counted, and only the frequent ones are kept as features. The most frequently occurred 'Nouns' and 'Noun phrases' usually refers to aspects or features in our consideration. The occurrences of feature are calculate as:

$\sum$ freq (f_i)…… where f_i *is a feature*

5. *Feature Clustering:* To uniquely identify the feature here we propose the ***clustering*** method to be used as feature classification approach. The features like "***Screen***" and "Display" would have the same meaning when we consider mobile products. The features are same or either called as synonyms. These conflicts then further may lead to misinterpretation of feature ranking. So to overcome this conflict a clustering algorithm can be applied. We propose synonym clustering to obtain unique features.

6. *Sentiment Classification on Features:* The Pros and Cons reviews can be categorized positive and negative opinions on the feature. These reviews are valuable training samples for learning a sentiment classification. Thus the ***Pros*** and Cons reviews used to train a sentiment classifier, which is in turn used to determine consumer opinions (positive or negative) on the aspects in free text Reviews.

7. *Probabilistic Feature Ranking:* By using probabilistic feature ranking algorithm we will identify the important features of a product from customer reviews. Generally, ***important*** features have the following characteristics: (a) they are frequently commented in ***customer's*** reviews; and (b) customer's reviews on these features greatly influence their overall opinions on the product. The overall opinion in a review is an aggregation of the opinions given to specific feature in the review, and various features have different contributions in the aggregation**.**

ISSN: 2321-8169
3894 - 3900

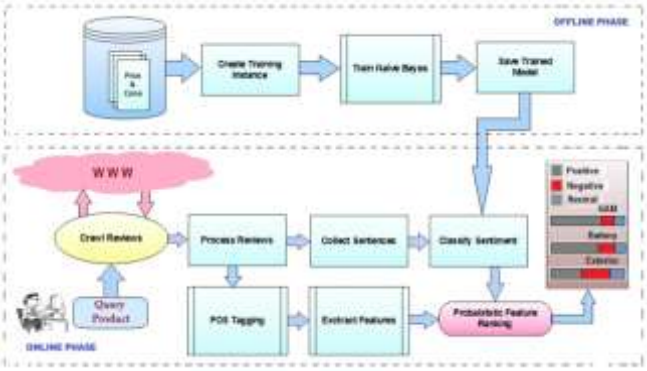

Figure 2. System Architecture

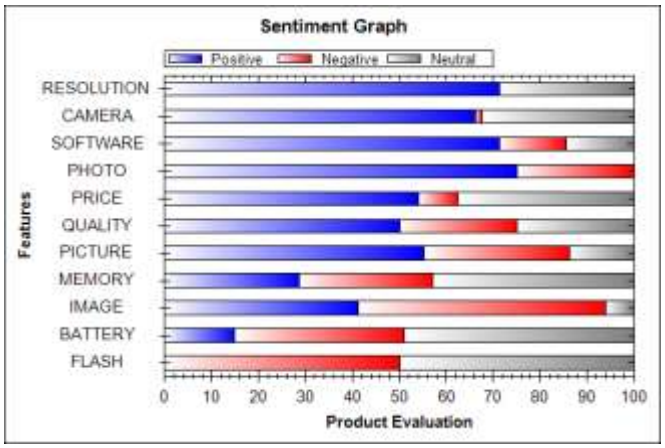

Figure 3. Output Design

IV. OUTPUT DESIGN AND DATASET

Most of the work in sentiment analysis uses product reviews data for experiments. Mobile review data sets are available at (http:// www.cs.cornell.edu/People/pabo/mobile-review-data). Many other datasets are available online as multi-domain sentiment (MDS) dataset. (http:// www.cs.jhu.edu/mdredze/datasets/sentiment).

V. RESULT ANALYSIS

The proposed Feature Extraction and Sentiment Labeling techniques are quite impressive when we analyzed the results on different queries. We firstly extracted the product features i.e Noun/ Noun phrases through Stanford Parser[1] and then collected the frequently occurred features. The precision and recall of the results is shown table 3. The resulting graph is plotted based on collected points from table 3.

Query Product	#Reviews	#Features	#Retrieved Features	#Correctly Identified Features	Recall	Precision
Blackberry Z10	52	79	62	54	0.78	0.87
Camera	38	46	36	30	0.78	0.83
Moto-G	45	60	46	37	0.77	0.81
Printer	34	49	37	28	0.76	0.75

Table 3. Feature Extraction Results

Query Product	#Reviews	#Sentences	#Correctly Identified Sentences	#Correctly Labeled Sentences	Recall	Precision
Blackberry Z10	52	436	370	355	0.85	0.88
Camera	38	235	191	178	0.81	0.83
Moto-G	45	268	209	191	0.78	0.79
Printer	34	210	158	119	0.75	0.79

Table 4. Sentiment Labeling Results

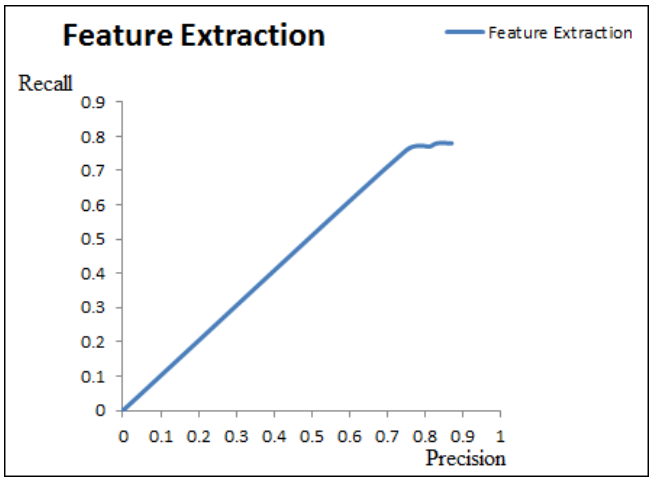

Figure 4. Precision Vs. Recall for Feature Extraction

Further we analyzed the results on sentiment labeling with respected to sentences extracted from free text reviews. Total 928 sentences were tested against the proposed Naïve Bayes classifier and the accuracy in terms of precision is calculated. We found the accuracy of Naïve Bayes around 82%, which was much effective in ranking the features. The precision and recall calculated are shown in table 4, and the resulting graph of Precision Vs. Recall is plotted in figure 5.

Moreover the resulting graphs were analyzed to see the effectiveness of the systems. As the ideal system follows the diagonal of the Precision Vs. Recall graph, we compared the lines which intersects the points on the graph with the diagonal and the resulting lines. The sentiment labeling proceeds in directly proportion whereas the feature extraction follows the diagonal upto 0.8 and then proceeds horizontally.

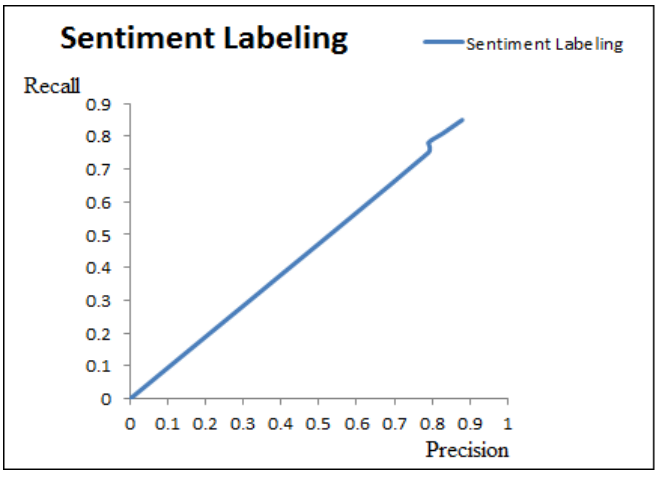

Figure 5. Precision Vs. Recall for Sentiment Labeling

1. http://nlp.stanford.edu/software/lex-parser.shtm

IV. CONCLUSION

In order to conclude the proposed product feature ranking framework, we crawled product reviews from e-commerce forum websites, such as CNet.com, Viewpoints.com Amazon.com etc. This corpus is available by request for future research on aspect ranking and related topics. Product feature ranking is beneficial to a wide range of real-world applications. We investigate its usefulness in two applications, i.e. document-level sentiment classification that aims to determine a review document as expressing a positive or negative overall opinion, and extractive review summarization which aims to summarize consumer reviews by selecting informative review sentences. We conducted various experiments to evaluate the effectiveness of feature ranking in these two applications and achieve significant performance improvements.

ACKNOWLEDGMENT

I express great many thanks to Prof. S. S. Bere and Department Staff for their great effort of supervising and leading me to accomplish this fine work. They were a great source of support and encouragement. To my friends and family, for their warm, kind encourages and loves. To every person who gave me something too light along my pathway. I thanks for believing in me.

REFERENCES

[1] Zheng-Jun Zha, Member, Ieee, Jianxing Yu, Jinhui Tang, Member, IEEE, Meng Wang, Member, IEEE, And Tat-Seng Chua,"Product Aspect Ranking And Its Applications", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014

[2] A. Ghose and P. G. Ipeirotis, “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics,” IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 1498–1512. Sept. 2010.

[3] Ayesha Rashid, Naveed Anwer, Dr. Muddaser Iqbal, Dr. Muhammad Sher,"A Survey Paper: Areas, Techniques and Challenges of Opinion Mining",IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013

[4] Neha S. Joshi, Suhasini A. Itkat,"A Survey on Feature Level Sentiment Analysis",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 014, 5422-5425

[5] M.Govindarajan ,Romina M,"A Survey of Classification Methods and Applications for Sentiment Analysis ",The International Journal Of Engineering And Science (IJES) ||Volume||2 ||Issue|| 12||Pages|| 11-15||2013|| ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805.

[6] Bo Pang and Lillian Lee,"Opinion Mining and Sentiment Analysis.”

[7] Bing Xiang,Liang Zhou,"Improving Twitter Sentiment Analysis with Topic-Based Mixture Modeling and Semi-

Supervised Training",Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 434–439,Baltimore, Maryland, USA, June 23-25 2014.

[8] M Fan, G WU "Opinion Summarization of Customer comments" International conference on Applied Physics and Industrial Engineering in 2012.

[9] N, Anwer and A, Rashid "Feature Based Opinion Mining of Online Free Format Customer Reviews Using Frequency Distribution and Bayesian Statistics" Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on 16-18 Aug. 2010.

[10] N. M. Shelke, S. Deshpande and V. Thakre "Survey of Techniques for Opinion Mining" International Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012.

[11] Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima," Sentiment Analysis Based Approaches for Understanding User Context in Web Content", 978- 0-7695-4958-3/13, 2013 IEEE.

[12] MC Wu, YF Lo and SH Hsu, "A fuzzy CBR technique for generating product ideas" Expert Systems with Applications, 34 (1), pg. 530-540 January 2008.

[13] Turney, Peter D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised, Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, Pennsylvania, USA, July 8-10, 2002. Pp 417-424. NRC 44946.

[14] Keisuke Mizumoto, Hidekazu Yanagimoto and Michifumi Yoshioka, "Sentiment Analysis of Stock Market News with Semi-supervised Learning", IEEE Computer Society,IEEE/ACIS 11th International Conference on Computer and Information Science, p.325-328,2012.

[15] Kim S, Hovy E. Determining the sentiment of opinions. In: Proceedings of interntional conference on Computational Linguistics (COLING'04); 2004.

[16] Segaran, T. (2007), Programming Collective Intelligence. Sebastopol: O'Reilly Media, Inc.

[17] Dave K., Lawrence, S. & Pennock, D.M. (2003), "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews". In Proceedings of the 12th International Conference on World Wide Web, p. 519- 528.