IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. VI (Jul – Aug. 2014), PP 78-85
www.iosrjournals.org

Web Logs Analysis for Finding Brand Status

Aditi B. Raut[1], Radha Shankarmani[2]
[1]*ME-Student, Computer Engineering, S.P.I.T, Mumbai University, India*
[2]*Information Technology, S.P.I.T., Mumbai University, India*

Abstract: *Due to rapid development of the web there has been vast increase in the user generated contents available in the form of blogs ,product reviews sites, web- forums and online social networks etc. Such reviews are very useful to the companies, as they contain valuable information about what aspects of the product, are driving the sales up or down.But most of the available content is in the form of unstructured data from where extraction of information is a challenging task. In order to find potential risk, it is essential for companies to gather and analyze information about competitor products and strategies. Companies use opinion mining techniques to track customer's response, in order to efficiently market their products, identify new opportunities, weaknesses and manage their reputations. In this paper, brand status model is proposed based on the various parameters which affect brands reputation. Also, the products belonging to different brands will be compared on the basis of predefined aspects. This model uses aspect based sentiment analysis technique to find sentiments about product's features and brand.*

Keywords: *Brand comparison, Opinion orientation, Product aspects, Product reviews, Sentiment analysis*

I. Introduction

In today's world, word of mouth is the primary factor behind 20 to 50 per cent of all purchasing decisions. Product reviews are posted online and opinions are widely spread through social networks. Web sites or blogs are also used, by many users, as a way to express their views about the brands. Due to sheer volume of information available today consumers prefer to make purchasing decisions, largely independent on the product information provided about the companies. The right messages are passed on, which expand within the concerned networks, affecting brand perceptions, share market.

In recent years, a number of online shopping customers and merchants are highly increased as a result of the exponential development of World Wide Web and e-commerce. Companies can utilize such online textual content in an effort to gain insight into consumers' opinions regarding available products and services. Ignoring consumer generated sentiments might put companies in a competitive disadvantage and could also create significant brand image problems. To know the customer requirements, product manufacturers encourage customers to share their views in the form of reviews on the products or services they have used in internet forums, discussion groups, and blogs etc. The customers can give opinion about product in the form of reviews at merchant sites, e.g. amazon.com, engadget.com, cnet.com, and epinions.com. If extracted and summarized, such opinions could provide valuable data for decision makers.

Usually, customers go through online reviews of a product before purchasing it. From product manufacturer viewpoint, it is most important to comprehend the customer's preference for product development, promotion and marketing. Using this information manufacturer can find weakness and strength of their product to improve it. Also, it is important to consider what aspects of products are liked and disliked by customers, because such online word of mouth actions provide new and valuable sources of information for business intelligence. Many businesses are now using opinion mining techniques in order to track customer responses to work accordingly. Sentiment Analysis is the task of retrieving and classifying opinions about certain products or its aspect as a positive or negative [1].

The process of opinion summarization is three steps procedure. In first step to perform product aspect identification in online reviews, next step is opinion words extraction associates with particular aspect. Finally generate a valuable summery, which shows aspect based product comparison.

This paper covers some techniques on extracting and summarizing opinions using aspect based opinion mining. The reviews are extracted from different sites and blogs after that preprocessing is done in order to remove noise. Further we apply methods to extract opinion and find aspect polarity. Here, we use different methods depending on the type of review format to extract opinionated text to find its polarity with the help of predefined aspects. For example Pros and cons, free text reviews etc.The aim of this paper is to find brand status by considering various parameters and its product opinions. Aspect based product comparison is also done. To perform this task data mining, natural language processing techniques are used.

II. Related Work

Mohamed M. Mostafa, this paper uses twitter data to analyze consumer's sentiment towards well-known brands such as Nokia, T-Mobile, IBM, KLM and DHL. To perform this task predefined lexicon of adjective is used with known polarity[2].Collecting and comparing consumer opinions of competing products from the Internet for business intelligence and for product improvement is an important issue.

There are several techniques exists to perform opinion mining tasks.

Gamgarn Somprasertsri et al. proposed an approach for extracting product features and associated opinion based on syntactic and semantic relation. They also use Product Ontology to find out similar features with different names. But this technique is not suitable for complex sentences [3].

Hu and Liu represented an innovative technique that uses association rule mining based on an a priori algorithm to perform extraction and summarization of customer reviews. The system that Hu and Liu developed is basically an unsupervised item set mining, works on frequently words representing aspects/features [4].

To improve the work over Hu et al, Liu et al [5] proposed an improved version of the original system based on language pattern extraction to identify product aspects/features from pros and cons type of reviews. They also attempt to extract implicit features.

Wenhao Zhang et al. developed a system called weakness finder to analyse consumer's sentiments in Chinese language online texts. The system uses aspect based sentiment analysis technique to help manufacturers to find their product weakness from Chinese reviews. The system considers both explicit and implicit product aspects. Then find associated opinion with their sentiment's polarity [1].

Zheng-Jun Zha et.al[6]proposed aspect ranking framework to identify important aspects of products from product reviews and the aspect ranking algorithm is developed which considers importance of aspects using aspect frequency and the influence of consumer opinions given to each aspect over their overall opinions.

Amanai K Samha et.al [7]proposed a framework to generate aspect based opinion summarization from customer reviews, using natural language processing, data mining and ontologies. The framework considers all possible product aspects.

Mita K. Dalal et.al [8] proposed a method to generate comparative feature-based summary using semi-supervised approach for mining online user reviews that can guide a user in decision making before online purchase.

Ahmad Kamal [9]illustrate, used both supervised machine learning and rule-based approaches simultaneously for mining possible feature-opinion pairs from subjective review sentences. The first phase describes classification of subjective.

III. Proposed System

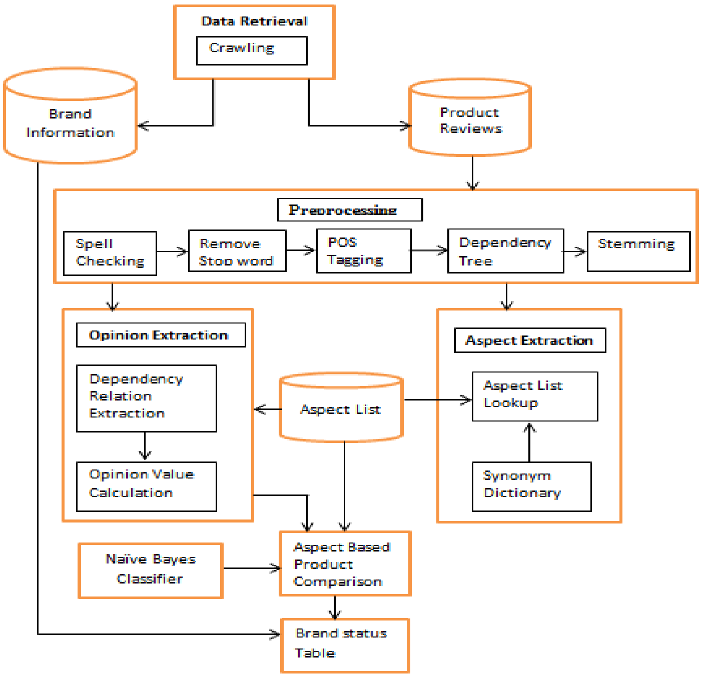


Fig. 3.1: System Architecture

3.1 Data retrieval:

In this phase data is extracted from various sources such as reviews sites cNet.com, amazon.com, egadget.com etc. and blogs. Retrieval of reviews is performed using crawler.

One of the most important parts of the proposed system is the Web crawler. A Web Crawler (also known as a Web spider or Web robot) is a program or automated script that browses the World Wide Web in a methodical, automated manner. The search term entered by the user is given as an input query to the Google Custom Search API which returns the results as a JSON object. As the crawler visits the URLs it identifies all the hyperlinks in the page and adds them to a list of URLs to visit, called the crawler frontier [10]. Here, we used Jsoup: java html parser API for extracting and manipulating data, using the best of DOM, CSS, and jquery methods. Here we use two databases namely product reviews and brand information. Data will be stored in appropriate database.

3.2 Preprocessing:

After crawling, extracted reviews from web pages are stored in the database according to its format.
The product reviews on the Web are in three formats:

There are three main review formats available. Different review formats may need different methods to perform the opinion extraction task.

Format (1)—pros and cons: The reviewer is asked to describe pros and cons separately.
Format (2)—pros, cons, and detailed review: the reviewer is asked to describe pros and cons separately and also write a detailed review.
Format (3)—free format: the reviewer can write freely, that is, no separation of pros and cons [11].

3.2.1 Aspect lookup table is generated using product specification provided by manufacturer and WorldNet dictionary which provides synonyms.

3.2.2 Spell checking:

Online reviews posted by users frequently contain spelling errors. Noise would be removed from reviews by performing spelling correction. In order to perform this task Jazzy spell checker is used. Also, some symbols and numbers are converted to its appropriate text form manually.

3.2.3 Stop word removal:

Firstly full text review is divided into sentences. Most of the English sentences include words like"a", "an", "of", "the", "I", "it", "you", and "and" etc. such words do not carry particular meaning. Information extraction from natural language can be done effectively and clearly by avoiding those words which occurs very often. We have created text file consists of stop words to remove such words from sentences replaces it with white spaces.

3.2.4 Part Of Speech Tagging:

The method of allocating different parts of speech tags such as noun, preposition, verb, adjective and adverb to a given text are known as Part-Of-Speech tagging (POS).The Stanford parser is used to generate the POS tagging of each word, present in the sentence. It is very essential as it helps us to find general language patterns.Opinion Mining requires adjectives and adverbs to find the aspect polarity. These can be obtained by using the POS Tagger.

3.2.5 Dependency parsing:

Customer reviews may have both subjective and objective sentences. Subjective sentences contain user's sentiment, emotion, conviction, rages, etc. Whereas, objective sentences usually holds factual information [12].

In this phase, sentences are first classified as subjective and objective sentences. As subjective sentences hold opinionated words, which intern helps in finding aspect based opinion polarity i.e. positive, negative. We first stored subjective sentences and then use Stanford dependency parser API to perform POS Tagging and dependency relation creation purpose.

3.2.6 Stemming:

Word stemming is the process of reducing transformed or derived words to their base or root form. For example, a stemming algorithm for English should stem the words using, used, use and uses, user to the root word, user. Here, porter's algorithm is used to perform stemming.

3.3 Aspect Extraction:

Aspect lookup table is createdto perform aspect based brand comparison, product aspects need to be extracted from product reviews. The aspects are as follows: Size, battery, price, camera, Sim cards, network, Weight, appearance, memory, processor, ease of use, applications etc.

Product reviews consists of implicit and explicit aspects.

If an aspect Aor any of its synonyms occurs in a sentence S*, A* is called an *explicit aspect*in S. If neither *A*nor any of its synonyms appear in Sbut *A*is implied, then *A*is called an *implicit aspect*in S.

Example:

- "image quality" in the following sentence is an explicit feature:

"The image quality of this phone is very nice".

- "This phone is heavy".

Weight is an implicit feature in the above sentence as it does not appear in the sentence [11].

for every aspect. Also, manually added domain dependent words which represent implicit aspects.

Example: price-{cost, reasonable, cheap, savings, expensive, affordable, etc.}
Memory-{storage, capacity, RAM, etc.}

3.4 Opinion Extraction:

Firstly, aspects are extracted from given text, after matching with predefined aspects. The next step is to find opinion associated with it. SentiwordNet lexicon is used to find aspect polarity. Methods would be used according to the format of the reviews.

Methods are as follows:

3.5 Pros and cons:

For pros and cons type of reviews, there is no need to find opinion orientations explicitly as it is already classified as positive and negative. In this case, only product aspects needed are identified using aspect lookup table. After that check whether the identified aspect is present in pros or cons reviews and mark it positive or negative and assign score accordingly.

3.6 Naive Bayes Classifier:

For detailed reviews naïve Bayes method would be used. Here input is considered as a blog text.Firstly, text needs to be preprocessed in order to remove noise. After preprocessing, the document will be converted into sentences. And algorithm would be used to classify reviews into positive or negative.

Naive Bayes classifiers are generally used for text classification. The Naive Bayes classifier is a probability classifier, based on Bayes' theorem.

In this case, the probability whether the document is positive or negative is calculated. Multinomial Naïve Bayes classifier is used for document classification.

It assumes that the probability of one word appearing in a document is independent of the probability of another word appearing. Firstly, sentences need to be extracted from the document, after that opinion will be extracted at sentence level.Here, pros and cons reviews are used in order to train the classifier, wherein the reviews are already classified into positive and negative. Since, manually tagging the training data is a time-consuming task.

Input:
A document d
A fixed set of classes C= {c1, c2…, cn}

Output:
A predicted class c ϵ C

Bayes' Rule for document d and class c:

$$P(c/d) = \frac{P(d/c)P(c)}{P(d)} \tag{1}$$

Classifier is trained to estimate prior probabilities for class c.

$$P(c) = \frac{Nc}{N} \tag{2}$$

Where N is that total number of documents and Nc is the number of documents that belong to class c.

The classifier also estimates the conditional probabilities that a term w appears in class c :

$$P(w,c) = \frac{count\ (w,c)+1}{count\ (c)+|v|} \tag{3}$$

Where count(w,c) is the total number of times term w appears in all the documents that belong to class c, count(c) is the number of terms in all the documents that belong to class c and |V| is the size of the vocabulary. The conditional probabilities using Laplace smoothing to avoid zeros.

Test document is classified using this formula,

$$C = {}^{argmax}_{\ } \ {}^{P(cj)}_{cj\,\in C}\prod\nolimits_{i\ \in\ positions}\ P(wi/cj) \tag{4}$$

3.7 Dependency Relation Extraction:

Extracting product aspects and its associated opinion from free text reviews is a challenging task. Here input is considered as detailed reviews.

Stanford dependency parser is applied to the sentences in order to extract appropriate dependency relation. This method is called as a rule-based system, and accepts subjective POS tagged review sentences as input along with dependency relationships information between words.

Generally, opinion words and product aspects are dependent of each other directly or indirectly using some semantic relations, first each sentence is tagged with its corresponding Parts of speech. After that it is converted into dependency tree using Stanford Parser. The dependency tree, also called as word-word relationship, encodes the grammatical relations between every pair of words [13].

Noun and noun phrases relate to product aspects, adjective represent opinions, and adverbs are used as modifier to represent the degree of opinion expressiveness [13].Therefore, it is defined as<*A, M, Op*>where, *A* is a noun phrase and *O* is adjective word possibly representing product feature and opinion respectively. M *represents* adverb that act as modifier to measure the strength of the sentiment used for opinion *O. M* is also used to capture the negative opinions explicitly expressed in the review. To obtain such a triplet, it is required to extract such relations which contain nouns,adjectives and adverbs etc.

The information component extraction mechanism is implemented as a rule-based system which analyzes dependency tree to extract information.

3.6.1 Algorithm for Information Extraction:

Input: DT- Dependency tree, ST-stop word list
Output: EI-Extracted Information <A, M, OP>

```
EI ←Ø
For each DT do
If ∃ at least one relation nsubj(w1, w2) in DT then
For each relation nsubj (w1, w2) ∈ DT do
If pos(w1) =JJ*&& pos(w2)=NN* && (w1,w2) ∉ ST then
 Assign A←w2, Op←W1
If ∃ advmod(w1,w3) ∈ DT && w3 ∉ ST then
 Assign M←w3.
End if
EI=EI ∪ {<A, M, OP>}.
Else if pos(w1)=VB* then
If ∃ a relation acomp(w1,w3) ∈ DT && w3 ∉ ST then
 Assign OP←w3
End if
If ∃ advmod(w3,w4) ∈ DT && w4 ∉ ST then
 Assign M←w4.
End if
EI=EI ∪ {<A, M, OP>}.
End if
End if
End for
```

```
    End if
End for
Return EI
```

```
For each DT do
    If ∃ at least one relation amod(w1,w2) || nn(w1,w2) in DT then
        If pos(w1)=NN* && pos(w2)=NN* && (w1,w2) ∉ ST then
            For each relation amod(w1,w3) ∈ DT &&pos(w3)=JJ* && w3 ∉ ST do
                Assign A←w1, w2 && op ←w3
            End for
            If ∃ advmod(w3,w4) ∈ DT && w4 ∉ ST then
                Assign M←w4
            End if
            EI=EI ∪ {<A, M, OP>}.
        End if
    End if
End for
Return EI
```

```
For each DT do
    If ∃ at least one relation amod(w1,w2) in DT then
        If pos(w1)=NN* && pos(w2)=JJ* && (w1,w2) ∉ ST then
            Assign A=w1 && op =w2
            For each relation conj(w1,w3) ∈ DT && pos(w3)=NN* && w3 ∉ ST do
                Assign A←w3 && op ←w2
            End for
            If ∃ advmod(w2,w4) ∈ DT && w4 ∉ ST then
                Assign M←w4
            End if
            EI=EI ∪ {<A, M, OP>}.
        End if
    End if
End for
Return EI
```

```
For each DT do
    If ∃ at least one relation nn(w1,w2) in DT then
        If pos(w1)=NN* && pos(w2)=NN* && (w1,w2) ∉ ST then
            For each relation nsub(w3,w1) ∈ DT && w3 ∉ ST do
                Assign A=w1, w2 && op =w3&& M=" "
            End for
            EI=EI ∪ {<A, M, OP>}.
        End if
    End if
End for
Return EI
```

The negation word related with the opinion word,changes the opinion polarity opposite to the one expressed in the sentence. The available negative words appeared in the review sentences are 'no', 'never' and 'not'. These negations can be extracted by capturing neg() relation between the opinion word.

Example:

- "The screen is not good".

nsubj (good-6, screen-3) neg (good-6, not-5)
The extracted feature, opinion pair is: (screen, good).

3.7 Opinion value calculation:

Once the triplet is extracted from reviews, noun (A) would be matched with the predefined aspects, existing in the database and if match is found then polarity of corresponding adjective is extracted and its sentiment score is generated using opinion lexicon, called SentiWordnet.

3.8 Brand Status Table

Here, information related to different mobile brands will be collected or gathered from authentic sources. This information would be used to find values for various parameters. Primarily, parameters like sales, rank, overall product rating, number of core products, successful products, sentiment, strength, passion, etc. are considered to find the brand status.

Table 3.1: Brand Comparison Table

Brand name	Rank	Sales (increase/decrease)	Sentiment (Positive: Negative)	passion	strength	Number of core models	Successful model	Overall Rating (5)	Status (growing/ declining/ stable)
Samsung	2	increase	19:1	55%	31%	3	3	4.5	Growing
Nokia	8	decrease	18:1	41%	36%	3	3	3.5	Declining

IV. Experimental Results

In this section, Experimental results of the proposed system areshown. Initially,Product reviews are crawled from different websites and aspect based opinion mining is performed.Pros and cons, Naive Bayes methodsare used to perform aspect based opinion mining. Following results are found after performing aspect based product comparison on different mobile model.

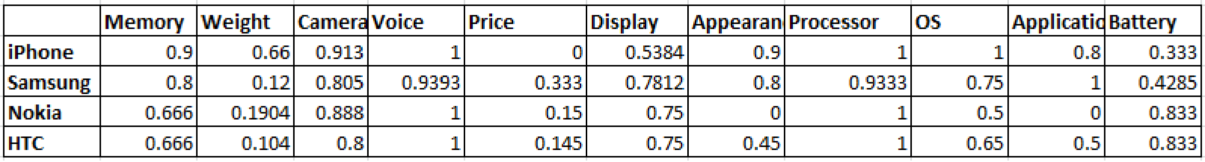
Here we consider, one mobile model from each brand namely iPhone, Samsung, Nokia and HTC etc. and reviews sites namely epinion.com,engadget.com,snapdeal.com and epinion.com etc.

The result table is as follows:

Gathered reviews for each model are not equal amount. Hence, normalization is performed using following formula

T=Total number of reviews for given product
A=Total number of reviews for given Aspect.
N=Negative reviews for Aspect.

$$Score = (A - T)/A \qquad (5)$$



	Memory	Weight	Camera	Voice	Price	Display	Appearan	Processor	OS	Applicatic	Battery
iPhone	0.9	0.66	0.913	1	0	0.5384	0.9	1	1	0.8	0.333
Samsung	0.8	0.12	0.805	0.9393	0.333	0.7812	0.8	0.9333	0.75	1	0.4285
Nokia	0.666	0.1904	0.888	1	0.15	0.75	0	1	0.5	0	0.833
HTC	0.666	0.104	0.8	1	0.145	0.75	0.45	1	0.65	0.5	0.833

Fig. 4.1: Aspect based product comparison table

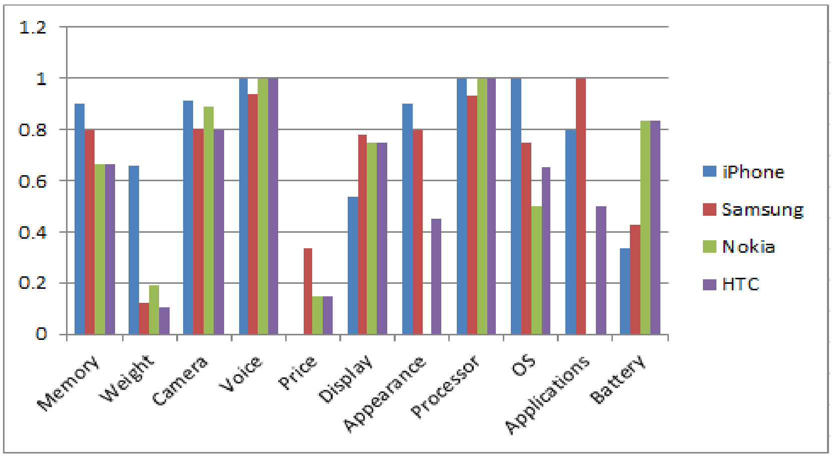


Fig. 4.2: Product Comparison Graph

V. Conclusion

In this paper, a system is proposed which determines the brand status, by considering different parameters and various sources like web logs and product reviews sites. Initially reviews are crawled from specific websites and brand related information is extracted from authentic sites. After that preprocessing is done in order to perform opinion mining. Aspect look up table is made with the help of predefined aspects and lexical resource. Later various methods like naïve Bayes and Dependency parsing are used to find opinion on product aspects and their results are added. Methods are decided according to type of reviews also, aspect based product comparison is made.

REFERENCES

[1] Wenhao Zhang, Hua Xu, Wei Wan" Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis" *Expert Systems with Applications 39 (2012) 10283–1029.*

[2] Mohamed M. Mostafa , "More than words: Social networks' text mining for consumer brand sentiments"*Expert Systems with Applications 40 (2013) 4241–4251*

[3] Gamgarn Somprasertsri, Pattarachai Lalitrojwong "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization" *Journal of Universal Computer Science, vol. 16, no. 6 (2010), 938-955 submitted: 15/9/09, accepted: 4/3/10, appeared: 28/3/10*

[4] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews" *in Proc. of SIGKDD, pp. 168-177. Seattle, WA, USA, 2004.*

[5] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*, 2005, *pp. 342-351.*

[6] Zheng-Jun Zha, Jianxing Yu, Meng Wang, Tat-Seng Chua "Product Aspect Ranking and Its Applications" *IEEE Transactions On Knowledge And Data Engineering 2013.*

[7] Amani K Samha,Yuefeng Li and JinglanZang "Aspect-Based Opinion Extraction from Customer Reviews".

[8] Mita K. Dalal., Mukesh A. Zaveri "Semisupervised Learning Based Opinion Summarization and Classification for Online Product Reviews"*Applied Computational Intelligence and Soft Computing Volume 2013, Article ID 910706.*

[9] Ahamad Kamal, "Subjectivity Classification using Machine Learning Techniques for Mining Feature-Opinion Pairs from Web Opinion Sources".

[10] G.Vinodhini,RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey"*International Journal of Advanced Research in Computer Science and Software Engineering,Volume 2, Issue 6, June 2012.*

[11] Bing Liu,"Sentiment Analysis and Subjectivity"*Handbook of Natural Language Processing*, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

[12] Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens, "Automatic Sentiment Analysis in On-line Text"*Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria – June 2007.*

[13] Tanvir Ahmad, Mohammad Najmud Doja, "Ranking System for Opinion Mining of Features from Review Documents*" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.*

[14] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, Aijun An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain" *IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 4, April 2012*

[15] Aurangzeb khan, Baharum Baharudin, "Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs"*Int. J Comp Sci. Emerging Tec Vol-2 No 4 August, 2011*

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up: sentiment classification using machine learning techniques" *Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86.*

[17] Ravi kumar, K. Raghuveer,"Web User Opinion Analysis for Product Features Extraction and Opinion Summarization"*International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.4*, October 2012.

[18] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, "Sentiment classification of Internet restaurant reviews written in Cantonese" *Expert Systems with Applications 38 (2011).*

[19] V. S. Jagtap and K. Pawar, "Analysis of different approaches tosentence-level sentiment classification"*International Journal of Scientific Engineering and Technology, vol. 2, no. 3, pp. 164–170, 2013.*

[20] S. Mukherjee and P. Bhattacharyya, "Feature specific sentiment analysis for product reviews" *in 13th International Conference on Intelligent Text Processing and Computational Linguistics, ser. Lecture Notes in Computer Science, vol. 7181. Springer, 2012, pp. 475–487.*