REVIEW SPAM ANALYSIS USING NATURAL LANGUAGE PROCESSING

[1] Jyoti G.Biradar

School of Mathematics and Computing Sciences of Computer Science

Rani Channamma University, Belagavi -591156, India

[2]Basavaraj A. Goudannavar

School of Mathematics and Computing Sciences Department Department of Computer Science

Rani Channamma University, Belagavi-591156 Karnataka, Karnataka, India

ABSTRACT- Communication is a vital part of traditional marketing; it is changed to posting reviews in websites/blogs. Many E-commerce merchants and vendors provide a platform to express their views/opinions on particular product and service. Thus, before purchasing any product (or) service it helps the customer to take decisions and for the vendors to promote their product and increase its sales. But, all the reviews posted by the users (or) customers need not to be written with true intention. User might have posted his review to promote (or) demote the product, such users or customers are considered as spammers. Thus, need to propose a new effective approach to classify each and every review as spam or non-spam. In this paper, the work is proposed for detecting spam/fake reviews on products using Term Frequency-Inverse Document Frequency (TF-IDF) weightage.

Index Terms: ***Reviews, spam detection, fake reviews, spammers, Term Frequency-Inverse Document Frequency.***

I INTRODUCTION

Sharing reviews through websites has become an important source of customer's opinions. It has become a habit for purchasing any product first they go through the reviews regarding the products, company or service etc. Therefore, if number of positive responses are more than the customers likes to buy the product. If the reviews are in terms of negative then customers move towards other products. Thus good reviews affect the economical condition and frame of organizations. The other part of these framework spammers tries to generate fake reviews for promotions or demotions of brands and mislead the customers. The main disadvantage here is there is no filtering or control on writing reviews as anyone can login and write their reviews. These reviews become an important source of data for buyers to decide and purchase the products. Thus spammers have chance to participate and affect the reputation of the products of the company.

The term spam describes unwanted message delivered to a large number of users from any company or website. Spam can be in many forms such as unwanted political mails, education, health care, personal, finance, computers, automotive, adult content and more. With the emergence of popularity of E-commerce websites they also provide an option to share customer's opinion or review about their products. Spammers can be defined as the type of malicious users who damages the information presented by legitimate users and also in turn fetches risk to the security and privacy of social networks. Spam is identified with following scenario:

- Unwanted junk mails are arrived in the mail box
- Generates burden for communications service providers and business to filter email.
- Phishing with the information by tricking into different links or entering details with good offers and promotions.
- Phishes: Behavior of the users matches the normal user to steal personal information of other legal users.
- Fake Users: Users who impersonate the profiles of genuine users to send spam content to the friends of that user or other users in the network.
- Promoters: are the ones who send malicious links of advertisements or other promotional links to others so as to obtain their personal information.

Spam reviews can be categorized into two types. In first type of reviews, spammers try to mislead readers by expressing undeserving positive opinions for some selected products for promotions or by writing negative reviews for damaging the name of the products. The second type of reviews includes only advertisements. Usually spammers focus on few goods and their service that users of computer are likely to be interested and they select the grey or black goods from market. In other way, spam can be stated as illegal, not only because of advertising the goods, but of the goods and services offered through advertisements are also illegal. From either sides business or company, reviews have following advantages: First, increase in rates of sales and conversion. Secondly, understand the feelings of customers about their brands, likes and dislikes about the products, improvements in shopping experience. Third, monitor and improve service offered by the company. Fourth, increase the traffic to website. Fifth, increase the loyalty of the customers. Furthermore, opinions are important for management of reputation and brand perception of product.

Detecting review spam is a challenging task due to the openness of writing product review on company's websites, there is no large-scale ground truth labeled dataset available. Spammers usually pose by different names of users which makes harder to eradicate spam reviews completely. Spam reviews looks perfectly normal until comparing with other reviews of the same products to identify review comments not consistent with the latter. The efforts of additional comparisons by the users make the detection task tedious and non-trivial. In this paper, Review spam identification is an important component to identify review spam, the data set consists of about 60k reviews. We need to provide real and trustful review mining results. In this paper, we introduce our review spam identification based on calculating term frequencies and using similarity measure.

II RELATED WORK

Harsha Patil et.al [01] proposed a framework for product aspect ranking, the system automatically try to find out few important aspects of products collected from different online websites and their related consumer reviews. Custumer reviews of a product are written in textual format; the first step will be parsing the reviews using Natural Language Processor (NLP) for identifying the aspects of particular product ,then to perform sentiment analysis as the second step, sentiment classifier such as Naïve Bays or SVM are used to classify the reviews as positive and negative sentiments. After sentiment analysis the third step will be implementing probabilistic aspect ranking algorithm to state the significance of aspects by simultaneously considering using aspect frequency and the influence of costumer opinions given to each aspect over their overall opinions.

u et.al [03] have compared different methods on machine learning approaches namely SVM (Support Vector Machine) and Naïve Bayes (NB) along with an ANN-based classifier method based on context of document-level sentiment classification. In this comparison, the main contributions of the authors are: (i) a comparison of two classifiers SVM and NB both the classifiers are dominant and a computationally efficient approach with an ANN-based approach under the same context; (ii) Using the ratio of positive and negative reviews a comparison involving realistic is carried out;(iii) a performance evaluation of ANN on a full version of the benchmark dataset of Movies reviews.

Ott et.al [02] used singleton review spam for the process of collecting the data. To achieve this they have prepared an artificial singleton review spam for the purpose of evaluating the proposed method and the singleton review spam are collected from real- world dataset ,temporal features are excluded.

Sushant Kokate et.al [07] proposed a behavioral approach for detecting the review spammers on rating for some targeted products. They have modeled an aggregated behavior scoring methods for rank reviewers to the degree that they express spamming behaviors. They have carried out experiment of proposed method for Amazon dataset for reviews of products manufactured by different companies. In order to detect the opinion spam, a centric spam detection method is used and concentrated on feedback data.

Wang et.al [04], proposed a system for detecting the hotels suspected reviews involved in spamming. For identifying the spammers they have made use of many features and designed two algorithms. They have designed a supervised method which results in list of hotels based on ranks which helps for human evaluation as it is difficult in suspecting the hotels which are involved in spamming.

Lim et.al [05], have considered the behaviors of reviewers by plotting social graph connecting different reviewers. In their work they have made a survey to discover the relationship of reinforcement between the a) reviewers trustiness b) reviews honesty and c) stores reliability for identifying the suspected spammers. The main drawback is that they do not handle single ton reviews as SRs includes adequate data on the graph to define their trustiness.

Mukherjee et.al [06] made the study on group spammers. They have proposed a three-step method for detecting group of spammers. They have mainly used frequent pattern mining algorithm as their first step for finding out groups of reviewers who frequently write the reviews on the targeted products together. As the next step, used feature extraction step for constructing the features for finding the group spammers. This method describes a novel approach for detecting a new form of spam activity. But lacks in addressing the singleton review spam problem, because only group of reviewers write reviews together at least 3 time s will be considered suspicious. In the same way the rule-based algorithm is designed by Ghose et.al [11] but failed during rule discovery step.

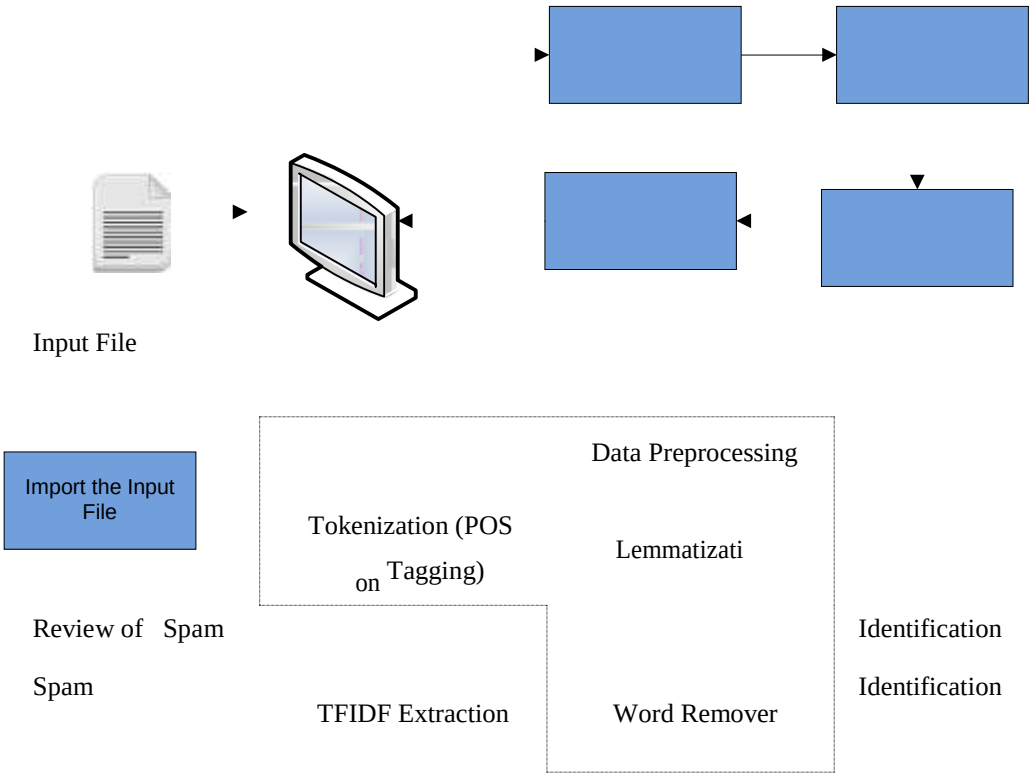
III METHODOLOGY

Given the consumer reviews of a product, first we identified the aspects in the reviews and then analyzed these reviews to find consumer opinions on the aspects via a sentiment analysis. The proposed review spam detection system mainly includes four main modules namely, Tokenization, where input text file is converted into words. Lemmatization, among separated words distinctive and relevant words are extracted. Word removals, common words which are of less significant are selected. Finally, TF-IDF “Term Frequency, Inverse Document Frequency" are used to measure the significance of each word in the document in terms of frequency of appearing across multiple documents. By measuring the similarity matrix, review spam is identified.

Tokenization

For any textual input the first step as pre-processing will be tokenization. A textual data input is the pattern of a collection of characters, punctuation, numbers, alpha-numeric etc. Therefore there is a need to convert graphic symbols into words for further processing. Certain character properties are used to classify strings into number of basic token types namely

- Word: sequences of letters;
- Number: sequences of digits;
- Separator: a Unicode separator, which includes punctuation characters, dashes and spaces.



Alphanumeric: a sequence of mixed letters, digits and symbols;

- Symbols: characters such as %, #, etc.;
- Control: characters such as line breaks, tabulations and control characters such as joiners.

In tokenization, a stream of text is converted into a stream of processing units called words or terms. There are three major goals of tokenization:

- first is to split the input text into tokens,
- second is to identify meaningful keywords and
- The third is to recognize sentence and word boundaries.

The output of tokenization are called as tokens which are in the form of words, phrases, keywords or symbols. Tokenization plays a significant role in lexical analysis.

Parts- of-Speech Tagging(POS)

Each word in a sentence defines its syntactic roles or parts of speech i.e., a word usage in the sentence. Part-of-speech is a process of assigning a word to its grammatical category. There are eight parts of speech in English: the verb, the noun, the pronoun, the adjective, the adverb, the preposition, the conjunction, and the interjection. In word processing part-of-speech (POS) taggers are used to classify words based on their parts of speech. Generating POS taggers helps with following reasons:

- Words like nouns and pronouns usually do not contain any sentiment. It is able to filter out such words with the help of a POS tagger;
- A POS tagger can also be used to distinguish words that can be used in different parts of speech.

POS taggers use a sequence of words as input, and provide a list of tuples as output, where each word is associated with the related tag. Part-of-speech tagging is what provides the contextual information that a lemmatize needs to choose the appropriate lemma.

Lemmatization

It is one of the important pre-processing steps. The task of Lemmatizers is to find distinctive and relevant words or candidate term using different statistical methods on word form or lemma level. The main concept of Lemmatizers is to map each token into its lexical headword or base word. This mapping transforms the word into its normalized form using POS tagging. A general lemmatize is made up of three parts: set of rules, lexicon of basic or normalized words and then finally lemmatization algorithm. These principles will set which words suffix need to be added or withdrawn to form a normalized form of a word. The rules will be in the form of if-then or if-then-else rules. Lemmatizers can be categorized into two major types: one is a manual approach that is based on handcrafted lexicon of lemmas and manually created affix rules. The second one is an automatic approach in which dictionary of lemmas and a set of affix rules are inferred from training data. For different languages, researches proposed different approaches of lemmatization. Limitations of the approach are: it is difficult to handle inflected natural languages have many words of the same normalized word, lemmatization of large dictionary is very time consuming strategy and lemmatization can be ambiguous as left can be normalized as left (adjective) or can normalize as leave(verb).

Word Removal

Few common words, which are of less significant or no content information in selecting the documents matching users need, is completely excluded from the vocabulary. These words are called as ‘stop words’ and the technique is termed as stop word removal or simple ‘Word Remove’. Stop words are those words which helps in formation of sentences or phrases. But, such words build a huge fraction of the text in documents, so in word processing, stop word filters are used in order to eliminate these words. Figure 2 shows the flowchart

data pre-processing steps for each document.

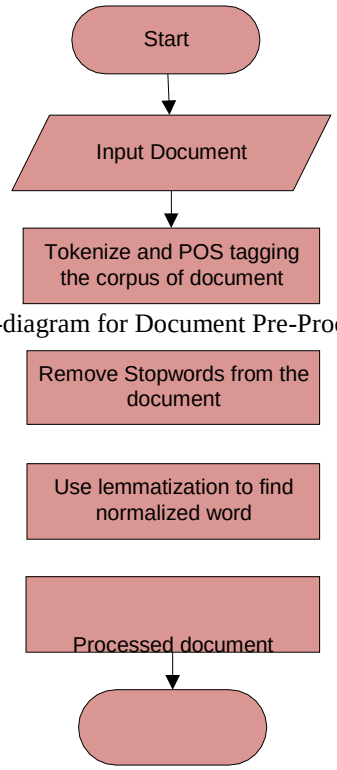


Figure 2: Flow-diagram for Document Pre-Processing.

For this purpose, a word of the list is built, this list can be provided by the user or system can automatically build it. Some schemes are also proposed for automatic generation of stop word lists. The stop word list contains articles such as 'the', 'a' and 'an', conjunctions like, 'and', 'or', 'but', and 'yet', prepositions like 'by', 'from', 'about', 'below', 'in' and 'onto' etc. This list also contains very frequent non-significant words which occur extremely often, but have little information content to make interesting decisions and this list have also words which occur rarely so have no significant statistical importance. Stop word list is contextual in nature or domain dependent, so according to application requirements this list can be customized.*F-IDF Extraction*

TF-IDF denotes "Term Frequency, Inverse Document Frequency" and used to measure the significance of each word in the document in terms of frequency of appearing across multiple documents. TF-IDF value is composed of two components: TF and IDF values which can be defined as follows:

a. *Term Frequency*: In the given document the term frequency (TF) defines the number of times a given term appears in that document. This frequency can be normalized to prevent the bias. Mathematically TF can be representing as the measure of the importance of a particular term ' ' within the particular '□'.

□□(□, □) = □□. □□. □□□□□□□□□□□□ □□ □ □□ □.

□□□□□□□□□□□□ □□(□, □) = □□. □□. (01)

□□□□□□□□□□□□ □□ □ □□ □

□□□□□ □□. □□. □□□□□ □□ □

b. *Inverse Document Frequency (IDF):* validates the term as rare or commonly occurs in the corpus. It is measured by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that quotient.

□□□(□, □) = |□□□| / |{□ ∈ □:□ ∈ □}| (02)

Where,

$|\boxtimes|$: Cardinality of *D*, i.e., total no.of documents in the corpus.

$|\{\boxtimes \in \boxtimes : \boxtimes \in \boxtimes\}|$: No.of. Documents where the term' $\boxtimes$ 'appears, i.e., $\boxtimes(\boxtimes, \boxtimes) \neq 0$. thus, if a particular term is not present in the entire corpus, it is divided by zero error. Therefore usually, the denominator is adjusted to $1 + |\{\boxtimes \in \boxtimes : \boxtimes \in \boxtimes\}|$

c. *TF-IDF:* A numerical statistic measurement that reflects the importance of a word in the document. It is often used as weighting factor in text mining and information retrieval. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. The tf-idf is calculated as :

$$\boxtimes\boxtimes * \boxtimes\boxtimes(\boxtimes, \boxtimes, \boxtimes) = \boxtimes\boxtimes(\boxtimes, \boxtimes)\boxtimes\boxtimes\boxtimes\boxtimes(\boxtimes, \boxtimes) \qquad (03)$$

d. A high weight in tf*idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the ratio inside the idf's log function is always greater than 1, the value of idf (and tf-idf) is greater than 0. As a term appears in more documents then ratio inside the log approaches 1 and making idf and tf-idf approaching 0. If a 1 is added to the denominator, a term that appears in all documents will have negative idf, and a term that occurs in all but one document will have an' idf 'equal to zero.

e. The two aspects neglect the frequency in the terms of a certain category. Because of the problems we have to add a weight to the original TF•IDF. The added weight considers the frequency of the term, which is in a particular category in the whole text collection, rather than simply consider the frequency of the term which is in the other documents of the whole text collection.

IV EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed approach, we have built an in-house data set of spam reviews and reviewers using human collected reviews from three different review websites Auto_parts_warehouse.com, Dhgate.com and neweggs.com websites, with different characteristics. We have randomly collected and selected reviews among these three review website. In order to detect spam and non-spam opinion reviews by applying data mining method, the data should first be preprocessed to get better input data for data mining techniques. The first step is Tokenization consists of separating strings into words called as tokens which are in the form of words, phrases, keywords or symbols. After tokenization, each token is labeled with Parts-of- speech tags .The next step is Stopwords removal and lemmatization which deletes tokens that are frequently used.

4.1 Similarity Measures:

Spam reviews are detected by the similarity measures between different reviews. The component accepts input as the feature matrix founded by calculating TF-IDF and finds percentage of matching of features from the database to identify the review as spam or non-spam. Similarity measure finds the similar words by mapping words to keywords or feature vector.

Let, $\boxtimes = \{\boxtimes_1, \boxtimes_2, \boxtimes_3, \ldots . \boxtimes_{\boxtimes}\}$ be the reviews with its features extracted from Auto_partswarehouse.com.and

$\boxtimes\boxtimes = \{\boxtimes\boxtimes_1, \boxtimes\boxtimes_2, \boxtimes\boxtimes_3, \ldots . \boxtimes\boxtimes_{\boxtimes}\}$ are the reviews with its features extracted from neweggs.com.

For detecting the spam and non-spam using similarity measure, the two feature matrix are compared to find the matching number of features or its equivalent synonyms between the reviews of both matrixes.

Conceptual level similarity measure between two review documents R and RN is defined as follows.

$$\square(\square, R\square) = \square_{\square} - \square_{\square}(\square, \square\square) \qquad (04)$$

Where,

$\square_{\square}$= Total number of feature

$\square$. $\square(\square, \square\square)$= Hamming distance between reviews$\square$, $\square\square$.

This similarities measured is used to classify the reviews as spam and non-spam based on threshold value T The classification rule is given by:

- If $\square(\square, \square\square)$ is in *[T, NC]*, then $\square$ and $\square\square$ are spam.
- Further for spam reviews $\square$ and $\square\square$ if $\square(\square, \square\square) = \square_{\square}$, then the reviews $\square$ and $\square\square$ are need to be checked.
- Similarly, for non-spam review $\square$ and $\square\square$ if $\square(\square, \square\square) = 0$, then the review $\square$ and $\square\square$ are unique, otherwise partially related.

Table 1. Similarity measure

Types of Reviews	Similarity Measures	
	Existing system	Proposed system
Spam Reviews	19.43%	17.81%
Non-Spam Reviews	80.57%	82.19%

Using the customer reviews from the database, in which every review is compared with all the other reviews in the dataset using similarity measure to find spam and the non-spam reviews. Table 1, shows the result of similarity measure for our proposed approach with comparing the existing approaches. Figure 3, shows the result of percentage of spam reviews obtained from the collected three different websites Auto_partswarehouse.com with 18.64%, Dhgate.com with 16.89% and neweggs.com with 14.69% respectively.

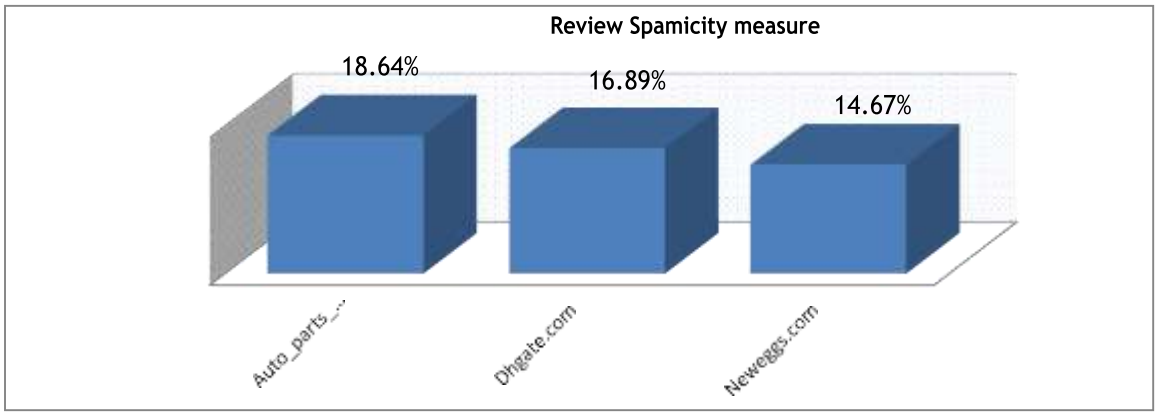


Figure 3. Result of Spamicity

V CONCLUSION

This proposed work is intended to detect trustworthiness of customer reviews. Defining and classifying a review into a spam and non-spam reviews is an Important and Challenging problem. . Many methods are developed and used by various researchers to discover review spamicity. Here the proposed methodology to find out the spamicity of reviews is done using Term frequency – Inverse document frequency and Similarity Measure approach. Proposed approach provides an enhanced summary of all spam and non-spam reviews for customers to easily decide whether to purchase the product or not and which are the reviews to be trusted while buying the product. Experimental outcome demonstrate the efficiency of the proposed technique in detecting spam and non-spam reviews.

REFERENCES

[1]. [1]. Harsha Patil, "Survey on Product Review Sentiment Analysis with Aspect Ranking" International Journal of Science and Research (IJSR), Vol. 4, No. 12, 2015.

[2]. [2]. M Ott , " Finding deceptive opinion spam by any stretch of the imagination", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 309 –319, 2011.

[3]. [3]. G Wu, "Merging multiple criteria to identify suspicious reviews", Proceedings of the fourth ACM conference on Recommender systems, pp. 241-244, 2010.

[4]. [4]. Wang G, " Identify online store review spammers via social review graph", ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 4, 2012.

[5]. [5]. Lim E-P, Nguyen V-A, Jindal N, Liu B, and Lauw H W. Detecting product review spammers using rating behaviours ,19th International Conference on Information and Knowledge Management and Co-located Workshops, pp. 939 - 948, 2010.

[6]. [6]. Mukherjee A, "Detecting group review spam", 2011.

[7]. [7]. Sushant Kokate, "Fake Review and Brand Spam Detection using J48 Classifier , (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 , No. 1, pp. 3523-3526, 2015.

[8]. [8]. J. C. Bezdek , "Convergence of alternating optimization," J. Neural Parallel Scientific Comput., Vol. 11, No. 4, pp. 351–368, 2003. [9]. X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in Proc. WSDM, New York, NY, USA, pp.

a. 231–240, 2008.

[9]. [10].G. Erkan ,"LexRank: Graph-based lexical centrality as salience in text summarization," J. Artif. Intell. Res, Vol. 22, No. 1, pp. 457– 479, 2004.

[10]. [11].O. Etzioni , "Unsupervised named-entity extraction from the web: An experimental study," J. Artif. Intell., Vol. 165, No. 1, pp. 91– 134, 2005.

[11]. [12].Ghose, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," IEEE Trans, Vol. 23, No. 10, pp. 1498 – 1512, 2010.

[12]. [13].V. Gupta, "A survey of text summarization extractive techniques," IEEE, Vol. 2, No. 3, pp. 258 – 268, 2010.

[13]. [14].W. Jin, "A novel lexicalized HMM-based learning framework for web opinion mining," in Proc. 26th Annu. ICML, pp. 465 – 472, 2009.

[14]. [15].M. Hu, "Mining and summarizing customer reviews," in Proc. SIGKDD, pp. 168 – 177, 2004.

[15]. [16].K. Jarvelin , "Cumulated gain-based evaluation of IR techniques," ACM Trans. Inform. Syst, Vol. 20, No. 4, pp. 422 – 446, 2002.