iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** II **Month of publication:** February 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Effective review selection using micro reviews and feature level extraction

Miss. A. Jenifaasheer. M.Sc, M.Phil[1], Dr.Anna Saro Vijendran. PhD[2]
Research Scholar, Associate Professor
Department Of Computer Science, SNR College of Arts and Science

Abstract: ***The online review about the product helps the user to decide the quality of product or service. The task of identifying appropriate review and distill useful information to take decision is very difficult. The complication arises with unregulated, lengthy and redundant description in reviews. Micro reviews, arising trend to extract useful information from reviews in social networking. Micro reviews are small, focused and compact (in 200 characters long). In this paper, we proposed a methodology by using micro review as an objective to extract set of review content about the entity effectively. Here two methodologies is used as matching the review content with micro review and selecting set of reviews by using feature based opinion mining. The reviews get classified by using naive bayes algorithm to produce effective result. The evaluation gets performed in the data collected from foursquare and yelp.***
Keywords: Micro-review, review selection, opinion mining.

I. INTRODUCTION

The web resources contain generous amount of review content. Though the information is useful the avalanche of online review faces many challenges. The user find difficult to find the required data from the overloaded content. The process of obtaining useful, detailed and authentic information about the product or place is tedious [2] [3]. With the advancement in social networking, the emanation of new type of online review content is recognized. The new type of content is termed as micro reviews, which helps the user to report about the service in current location or about the product. For example, at four square the user can sign in at local venues like restaurants, coffee shops. After check in they can leave their experience in small message as micro reviews about the place. The micro review is noted as tip in four square terminologies. In the case of restaurants, tips and opinions are frequently used terms. The tip (e.g., what to order), opinions (what is great or not), or actual “tips”. The required information based on user interest can be represented as micro reviews. Micro review refers to the concise opinion about an entity or topic often adhering to some constraints.

The advantages in micro review as:

A. The length is restricted, so they contain only distilled and related information.
B. The reviews are getting registered in the site after the users check in, so they can register the original expression which is spontaneous and unchanged.
C. As the comments are entered from the location which provides the authenticity.

The reviews and micro review are contrast to each other, as micro reviews are small, focused, within 200 character length. Review is lengthy, detailed which causes difficulty in the study of specific characteristics. By combining both aspects the important part is to develop the detailed review which focuses on aspects of a venue that are of true importance to users. Micro blogging supports authenticity by filtering the reviews depend on check in details. The problem considered is from the collected set of review and collected set of tips; want to separate the best envelope of tips. This will support the mobile applications effectively which able to display small amount of reviews. The problem in review selection causes the coverage problem [4], where the selected reviews want to pretense different facet of the item (e.g., product attributes), and the polarity of opinions about the item (positive and negative). The supervised and unsupervised techniques are commonly used to extract the aspects from the review. They are successful generally but cannot generalize to arbitrary domains and they suffer from the broadness of the topic definition. The tip provides a contract way to extract the aspect of the item and the sentiment of the user. The important review content which needed to elaborate can effectively choose by a tip. The selected reviews are compact and not get depart from the important points. The important perceive is mobile devices where the screen is small and time is short. In this paper micro review is studied and shows how they are effectively used to get the useful reviews. To mine the micro reviews and combines them with the full-reviews. The goal is to

maximize the efficiency and coverage by novel method. The proposed method includes integrated linear programming, formulation, greedy heuristics and opinion mining.

The rest of the paper is organized as follows, Section II explains the related work, Section III describes the Proposed System

II. RELATED WORKS

The mining technique deals with ability to choose best set of reviews. In [5], the goal is to choose best set of attributes from available attributes. To obtain all the positive and negative aspects of the attributes [6] present in the reviews. The problems which are get related with review selection [8] which get relates to quality, diversity and coverage. Lei Zhang, Bing Liu [9] proposed a method over Feature – Based Sentiment Analysis and also to prune Non –Opinated Features. In [10] an overview on feature based opinion mining and summarization is proposed. The porter stemmer algorithm is explained by Wahiba Ben Abdessalem Karaa in [7]. The shallow semantic analysis is performed in [11], the logics and sentiment are get classified and related to the sub sentence which are quite close to each other, topic and sentiment located in different sub-sentences, either being adjacent or not, but the different sub sentences are independent of each other, no parallel structures any more. In [12], the parsing dependency for opinion mining is used. The relation between head and determinants is used to define a structure by dependency grammar. The dependent is a modifier or complement and the head plays a more important role in determining the behaviors of the pair. The authors want to compromise between the information loss of the word level dependency in dependency parsing as it does not explicitly provide local structures and syntactic categories of phrases and the information gain in extracting long distance relations. Hence they extend the dependency tree node with phrases." Hu *et. al* [13] used frequent item sets to extract the most relevant features from a domain and pruned it to obtain a subset of features. They extract the nearby adjectives to a feature as an opinion word regarding that feature. Using a seed set of labeled Adjectives, which they manually develop for each domain; they further expand it using Word Net and use them to classify the extracted opinion words as positive or negative.

III. PROPOSED SYSTEM

The overview of the method explained with a restaurant example. The input for the process is a collection of reviews R and a collection of tips. The goal is to select effectively and profoundly a subset of reviews S R that covers the set of tips. The matching between review and tips can be performed by selection, which is determined when a review R ∈ R covers a tip t ∈ T. In matching small set of reviews is selected which covers as many tips as possible. The number of covered tips is referred as selection coverage. The selection efficiency, defined as ability to capture the principle that the selected set should not contain too many sentences that do not cover any tip.

A. Relation between review and micro review

The granularity gets different between review and micro review. The micro review is single pointed and short, while review is multi-dimensional. If the point made by the tip appears within the text of the review then review covers a tip. The review gets divided in to sentences, which are semantic units with granularity similar to that of the tips.

The review R can be viewed as set of sentences $\{s_1,\ldots,s_R\}$ and denote U_s as union of all review sentences from review R. The matching function can be defined as $F:U_s \times T \rightarrow \{0,1\}$, where for the sentence $s \in U_s$ and a tip $t \in T$, then

$$F(s,t) = \begin{cases} 1, \text{ if s and t are similar} \\ 0, \text{ otherwise} \end{cases}$$

The sentence and tip should be matched if they are having same meaning. The three criteria to make the matching decision as,

1) Syntactic similarity, if bag of words represent same meaning.
2) Semantic similarity, if same concept is discussed.
3) Sentiment similarity, if same opinion is presented (positive or negative).

B. Selection coverage and selection efficiency

If a sentence s and a tip t are matched, then we say that s covers t. Then review R covers a tip t if there is a sentence s 2 R that is matched to the tip t. Given the collection of reviews R and the collection of tips T, and the matching function F, for each review R the set of tips T_R that is covered by at least one sentence of review. This can be extended to collection of reviews as,

$$Cov(S)=\frac{|U\ R \in S\ T\ R|}{|T|}$$

The efficiency Eff(R) of the review R is defined as the fraction of "relevant" sentences in R represented as,

Eff(R) = $\frac{|Rt|}{|R|}$

IV. EXPERIMENTAL ANALYSIS

The small set of reviews results in effective methods. The optimization problem increases when data set increases. As Effmaxcoverage and Effsetcover problems (1) are get solved by the algorithms to provide optimal solution. The Effmaxcoverage optimally represented as Integer Linear Programming problem, for which there are known algorithms for deriving an optimal solution. However, the ILP formulation, while optimal, may not be tractable for cases where the number of reviews is very large (1). The greedy selection algorithm is used to solve the NP-hard problem. The sub modularity property of the coverage function, the greedy algorithm that always selects the review whose addition maximizes the coverage produces a solution with approximation ratio 1-$\frac{1}{e}$ for the MAXCOVERAGE problem, where is the base of the natural logarithm [23]. That is, the coverage of the greedy algorithm is at least a 1- $\frac{1}{e}$ fraction of the coverage of the optimal algorithm (1).

A. *Matching reviews and micro reviews*

1) Syntactic similarity: When the review and micro review shares the same keyword it said to have syntactic similarity. The review sentence and review associated with vectors s and t. The size of the vocabulary is determined by the dimensionality of vector(1). Each vector entry signifies the importance of the corresponding word. The degree of similarity between the sentence and the tip is then measured as the cosine similarity [20]. Therefore we have:

Synsim(s,t) = cosine(s,t)

2) Semantic similarity: When the same concept gets explained by both the review and micro review provides the semantic similarity. The difference in the usage of word get differs but indicates same meaning .Here the Latent Dirichlet Allocation (LDA) (2) approach is used. LDA associates each tip t with a probability distribution over the topics, which captures which topics are most important for t. A commonly used distance measure between two probability distributions is the Jensen-Shannon Divergence (JSD)[2].To measure the semantic similarity between a review sentence and a tip, Semsim(s,t)= 1-JSD(Θ_s, Θ_t)

3) *Sentiment similarity- opinion mining:* This includes few steps for extracting the opinion. The pool of opinion generated in each step constitutes the next process. The steps include as:

1) Preprocessing 2) First Level Feature Extraction 2) Second Level Feature Extraction 3) Third Level Feature Extraction 4) Classifier.

B. *Preprocessing*

The reviews are present in the raw form so some preprocessing steps need to get perform in order to remove the noises. Here stemming process is carried out for effective extraction process to improve the efficiency.

1) *Porter stemming:* Stemming is a technique to reduce the derived words in to base words or to root words. The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. The Porter algorithm differs from Lovins-type stemmers [7] in two major ways. The first difference is a significant reduction in the complexity of the rules associated with suffix removal. The second difference is the use of a single, unified approach to the handling of context. Porter uses a minimal length based on the number of consonant-vowel consonant strings (the *measure*) remains after removal of a suffix. This algorithm supports excellent trade-off between speed, readability, and accuracy. Stems using a set of rules, or transformations, applied in a succession of steps.

Porter Stemmer Steps

Step 1: Gets rid of plurals and -ed or -ing suffixes.
Step 2: Turns terminal y to i when there is another vowel in the stem.
Step 3: Maps double suffixes to single ones: -ization, -ational, etc.
Step 4: Deals with suffixes, -full, -ness etc.
Step 5: Takes off -ant, -ence, etc.
Step 6: Removes a final –e

A typical rule as can be given as,

(M>0) → formality →formal

(M>1) → adjustable→ adjust

This indicates that, when the non-zero measure (m) get satisfied for the resulting stem then FORMALITY is replaced by FORMAL. The multi component suffix can be getting removed in iterative stages in porter algorithm.

2) *Auxiliary List Preparation:* Auxiliary verbs are a subclass of verb that prototypically marks tense, aspect, mood or voice. Auxiliary word list comprises a series of word sets like articles, verbs, comparatives, conjunctions, decreases (e.g. "less"), increasers (e.g. "extra"), negations (e.g. "not") and pro-nouns. These words will constitute a main feed of the algorithm. The proposed approach utilizes this seed in the construction of all extraction patterns.

3) *Feature extraction:* Feature extraction is related is related with dimensionality reduction. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps.

The dependency tree provides connections between distant words, which are useful in extracting long distance relations. Formally, we define the dependency parsing with phrase nodes as phrase dependency parsing. A dependency relationship which is an asymmetric binary relationship holds between two phrases. One is called head, which is the central phrase in the relation. The other phrase is called dependent, which modifies the head Pruning [9] away local dependency relations by additional phrase structure information, phrase dependency parsing accelerates following processing of opinion relation extraction.

a) *First Level Filtered Seed Extraction:* Let S0 be the filtered seed from the set of seed word collected from the list of review S' = S∩V. The seed words which are not present in the extant work are get filtered out. The polarity of each seed word gets provided. The polarity disambiguation step is used to overcome the problem of change in polarity due to usage of words. The end of this process we have a pool of newly discovered opinion words along with their polarity.

b) *Second Level Conjunction-Based Extraction:* In this level the performance is depend on the sentiment consistency that applies in conjunct words (e.g. "lightweight and well-built device"). So the algorithm discovers the new opinion words from the sentiment consistency theories. In this level the positive and negative expressions are get extracted. The sentiments go through the polarity disambiguation process like the previous step. At the end of this process we have an extended list of opinion words S' U C where C is the list of opinion words extracted from this level.

c) *Third Level Double Propagation Extraction:* In this level the recently extracted sentimental words and features are used to extract new sentimental words and features. This process will carried [10] out till no new words get extracted so called as double propagation method. The theory of double propagation is followed to extract the opinions. The assumption is that opinion target and opinion words are get attached e.g., nice place. Here nice is directly attached with place. So this process is iterative. Based on double propagation method and using the current list of opinion words, able to identify new opinion targets. Using this set of opinion targets we are able to extract new opinion words following the same logic. The iteration goes for i times till it find the new opinions. The words and sentence are getting linked with each other by certain relation.

i) Direct relation: one word depends on other directly.

ii) Indirect relation: Two words get related to a third word or one word depends on the other word through other words.

At end of this step opinion words are discovered p_i. At end of this process a list obtained which consists of D= S' U C U P_i. At this step advantage of intra sentential and inter-sentential sentiment consistency is considered. The intra-sentential consistency suggests that if there are other opinion words in a sentence with known orientation, then, the newly found word will get the accumulated sentiment of these words. When there are no other known opinion words in the sentence, the inter-sentential assumption is applied. At the end of this process the pool of new opinion words and their orientation is obtained

d) *Classification:* The supervised machine learning method is used to classify the polarity present in the data set. By using previously hand training data the positive and negative of the sentence get classified.

i) *Naive Bayes Classifier:* The effective supervised machine learning approach is used for classification. Here the features collected in double propagation are used as trained data sets to match the attributes.

$$Pr(C|rv) = \Pi_{wi\ \varepsilon\ w}\ pr(app_w|c)$$

Where rv is the review under consideration, w is a feature extraction words pair that appears in the given document, *Pr*(appw|class) is the probability that a feature extraction words pair appears in a document of the given class in training data, and *bc* is an estimated class. The probabilistic models computed by the Naïve Bayes classifiers were sorted by log posterior odds on positive and negative orientations for the purpose of ranking, i.e. by a "score" computed as follows

$$Score = \log Pr(+|rv) - \log Pr(-|rv)$$

Volume 4 Issue II, February 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Where *rv* is the review under consideration, *Pr (+|rv)* is the probability of rv being a review of positive polarity, Pr*(-|rv)* analogously is the probability of the review being of negative polarity.

e) Performance Evaluation: The data set related with restaurant is considered for performance evaluation. The greedy algorithm reduces the computation time effectively to get a set of review effectively. By using the feature selection the initial seed is considered as ground for set of opinion words. . The set of results focuses on the impact of each step at opinion-word discovery. The double propagation helps to discover new words effectively. In this approach the opinion are extracted not only from the domain specific words but also from the extracted opinion. The sentimental analysis is performed in each step and the positive and negative opinions are effectively captured by opinion classification method. Effectiveness of method is determined by precision or recall values.

$$\text{Precision} = \frac{TP}{TP+FP}$$

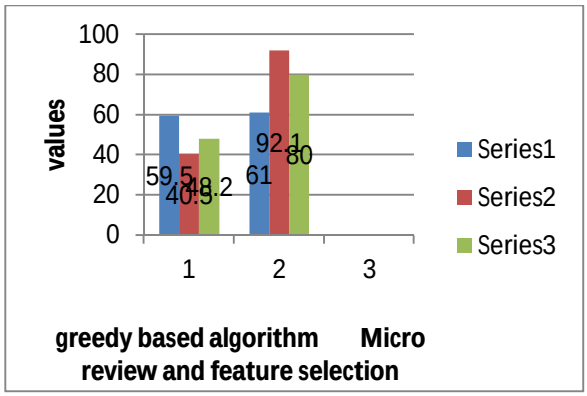
$$\text{Recall} = \frac{TP}{TP+FN}$$

i) True positives (TP) - number of reviews correctly labeled as belonging to particular class (positive/negative).

ii) False positives (FP) - number of reviews incorrectly labeled as belonging to particular class.

iii) False negatives (FN) - number of reviews were not labeled as belonging to the particular class but should have been labeled.

The graph plotted shows that the double precision naive bayes used in micro review and feature selection shows better performance when compared with the greedy algorithm.



Series 1- precision values
Series 2- recall values
Series 3- fmeasure values

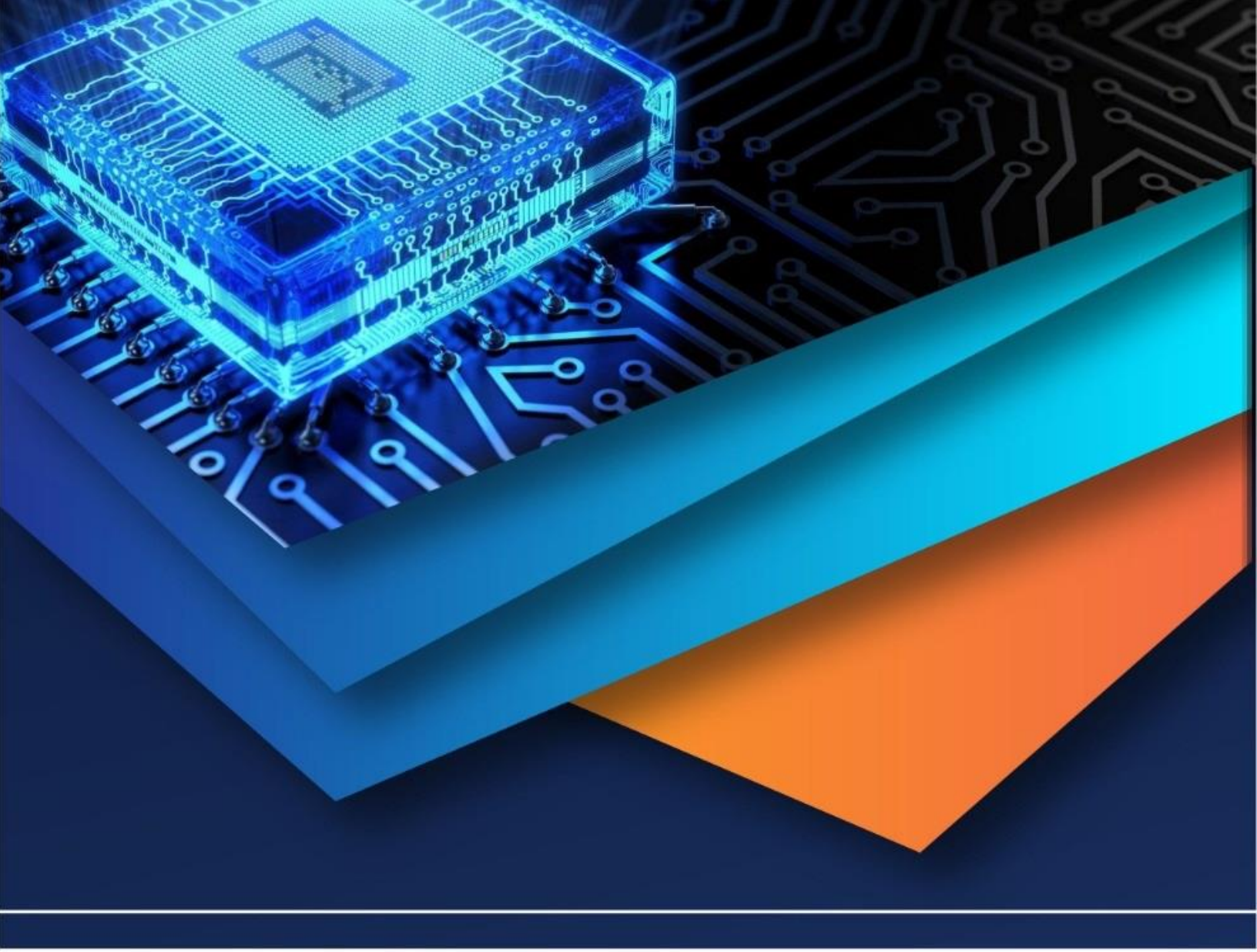
V. CONCLUSION

The micro reviews are presented as effective measure to cover a set of valuable reviews. The greedy algorithm is used to overcome the NP hard problem to improve the computation speed. In this paper an approach on the method for domain-specific opinion word discovery was presented. Word polarity is calculated automatically by following a set of polarity disambiguation procedures. The naïve Bayesian procedure is used to extract the features from the review. The Double propagation method is used to capture the association between features and sentiment words.

REFERENCES

[1] Thanh-Son Nguyen, Hady W. Lauw, Panayiotis Tsaparas, "Review Selection Using Micro-Reviews" in ieee transactions on knowledge and data engineering, vol. 27., no. 4, april 2015, pp.1041-4347.
[2] H. Gao, J. Tang, X. Hu, and H. Liu, "Exploring temporal effects for location recommendation on location-based social networks," in Proc. 7th ACM Conf. Recommender Syst., 2013,pp. 93–100.
[3] A. Ghose and P. G. Ipeirotis, "Designing novel review ranking systems: Predicting the usefulness and impact of reviews," in Proc. 9th Int. Conf. Electron. Commerce, 2007, pp. 303–310.
[4] Y. Lu, C. Zhai, and N. Sundaresan, "Rated aspect summarization of short comments," in Proc. 18th Int. Conf. World Wide Web, 2009,pp. 131–140.
[5] T. Lappas and D. Gunopulos, "Efficient confident search in large review corpora," in Proc. Eur. Conf. Mach. Learn. knowl. Discovery Databases: Part II, 2010, pp. 195–210.
[6] P. Tsaparas, A. Ntoulas, and E. Terzi, "Selecting a comprehensive set of reviews," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2011, pp. 168–176.

[7] Wahiba Ben Abdessalem Karaa, “a new stemmer to improve information retrieval”, in International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.4, July 2013.
[8] P. Sinha, S. Mehrotra, and R. Jain, “Summarization of personal photologs using multidimensional content and context,” in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.
[9] Lei Zhang, Bing Liu, “Identifying Noun Product Features that Imply Opinions”.
[10] Padmapani P. Tribhuvan, S.G. Bhirud, Amrapali P.Tribhuvan, “A Peer Review of Feature Based Opinion Mining and Summarization”, International Journal of Computer Science and Information Technologies, Vol 5(1), 2014, 247-250. ISSN: 0975-9646. Page No: 247-250.
[11] Chen Mosha,”Combining Dependency Parsing with Shallow Semantic Analysis for Chinese Opinion-Element Relation Identification”, IEEE, 2010, pp.299-305.
[12] Yuanbin Wu, Qi Zhang, Xuanjing Huang, Lide Wu,”Phrase Dependency Parsing for Opinion Mining”, EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, Volume 3
[13] Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, Xuanjing Huang,”Mining Product Reviews Based on Shallow Dependency Parsing”, SIGIR '09, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY