Research Journal of Applied Sciences, Engineering and Technology12(8):828-840, 2016
DOI:10.19026/rjaset.12.2783
ISSN:2040-7459; e-ISSN: 2040-7467

Submitted: April 17, 2015 Accepted: January 8, 2016 Published: April 15, 2016

Research Article

Effective Sentiment Analysis for Opinion Mining Using Artificial Bee Colony Optimization

T.M. Saravanan and A. Tamilarasi
Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu 638052, India

Abstract: Opinions play important role in the process of knowledge discovery or information retrieval and can be considered as a sub discipline of Data Mining. The huge quantity of information on web platforms put together feasible for exercise as data sources, in applications based on opinion mining and classification. An effective sentiment analysis process proposes in this research for mining and classifying the opinions. The phases of the proposed research are: (1) Data Pre-processing Phase (2) Potential Feature Extraction Phase (3) Opinion Extraction and Mining Phase and (4) Opinion Classification Phase. Initially, the datasets from various web documents get preprocessed and gives as part-of-speech tagged text. An Improved High Adjective Count (IHAC) Algorithm employs on the Part-Of-Speech tagged text to extract the potential features. Improved High Adjective Count Algorithm effectively optimizes the scores of the nouns to extract the potential features. An Artificial Bee Colony (ABC) Algorithm works under the IHAC algorithm for providing opinion scores and also for giving ranks for every noun. Max Opinion Score Algorithm can be then helpful to extract the opinion words followed by the classification phase, in which, ID3 algorithm utilizes to classify the review into three kinds positive, negative and neutral based on the opinions. The implementation is carried out on Customer Review Datasets and Additional Review Datasets with the aid of JAVA platform and also the experimentation results are analyzed.

Keywords: Artificial bee colony algorithm, ID3 algorithm, improved high adjective count algorithm, max opinion score algorithm, opinion mining

INTRODUCTION

The drastic development of World Wide Web has generated huge volume of data that engulf the domestic users of computers. The generated data has been contributed by internet users of anywhere in the world through their own thoughts or any information that they have observed or any commercial setups those want to engage in online business (Etzioni*et al*., 2005). Such kind of data makes extracting the vital and pertinent information, a challenging process. The cosmic nature of the data further reveals the bottleneck for its usage in business needs. Data mining techniques can be applied to extract information from such web data, which often referred as web mining techniques. Web mining discovers and extracts information from various web services and documents using data mining techniques (Chang *et al*., 2006). This novel research area can be said as interdisciplinary or multidisciplinary techniques, because this area is comprised and unified with other research areas such as data mining, text mining, databases, machine learning, multimedia, statistics, etc. There are three major operations formulate web mining techniques, namely, clustering (determines segments of users, pages, etc.), relationships (attempt to request URLs with association in some means) and sequential analysis (attempts to access the URLs with specified or logical order).

When dealing with most of the real-time problems, there will not be any precise boundaries; in fact, there will be overlapping boundaries in the clusters or associations obtained from web mining techniques. Moreover, web data may be partial and may have abundant exemplars (outliers) because of many factors that naturally occur while web browsing and logging. This necessitates the noise contaminated overlapping sets with exemplars, to be precisely modeled using web mining and personalization approaches. The web mining approaches should also be strong and fast enough to extract pertinent information from very huge internet database (Meena and Prabhakar, 2007). The internet database exhibit assorted nature by including organized and unorganized contents like images, audios and videos (Pang and Lee, 2008). The information in internet web are huge and hence, manifold of same information occur with dissimilarities in their word and

Corresponding Author: T.M. Saravanan, Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu 638052, India

formats representation. However, there is a mutual relationship between the important web information.

A consolidated web mining process is nothing but retrieving knowledge from the World Wide Web using traditional data mining methodologies and further uploading them as features of website (Hong, 2010). The web can be seen as a crucial messaging and marketing media. For instance, E-commerce websites are found to be a significant media for business. Here, data mining methods are need to be adopted for studying the user activities on the website. Generally, there are three categories of web mining methods. They are, web content mining, web structure mining and web usage mining. Web content mining uses search engines to extract knowledge from huge online databases and diverse collectionof websites, automatically (Chang *et al*., 2006). This process generates a structural report on a web page by finding complete set of hyperlinks from a given document. The web content mining strategies can be further classified into two strategic groups. Erstwhile, directly mines the document contents, whereas the latter strategy enhances the e-search engines further for content search. Web structure mining is a process of identifying the model behind the web link structures (Wonga and Lam, 2009). Here, graph theory is used to scrutinize the structure between the nodes and connection on a website.

Based on the category and data nature of the web structure, it can be classified into two types. They are:

- Pattern extraction from hyperlink on the net in which the hyperlink refers to web address in structural form that interconnects web page with other locations
- Mining document structure in which XHTML or HTML tags in the webpage can be investigated and represented using tree structures. In web usage mining, user patterns are identified and investigated by mining web associated data and log files (Zhai and Liu, 2006). The mining process helps to determine the patterns and their movements about the nature of the visitors, while browsing throughout the website. Retrieving the navigation patterns persists, due to the tracing and subsequent designing of browsing patterns and website structure, respectively. Frequency and duration of accessing a particular website mainly defines the users' nature in browsing that can be determined using log file, which records only information of web page sessions.

At certain circumstances, user prefers offline web pages on the internet for reasons such as limited download shots, offline data availability, data backup and many more (Hong, 2010). This necessitates the downloading of raw data from the internet as these are going as inputs to software to be used for data mining functions. Recent era has shown dramatic technological growth on products with minor deviations. Each product has to be tested systematically, where internet often plays significant role in analyzing the products by collecting required knowledge.

In this study, for feature extraction we will use opinion mining under consideration for extracting the reviews of the product from different web pages. Providing reviews corresponding to the ratings does not only help users gain more insight in to item's quality but it also helps to compare different items. This feature based extraction using opinion mining addresses this need by performing two main tasks. Feature identification which means extracting and identifying feature from the users reviews. Rating prediction L estimating the numerical rating of the feature of the product. The primary intension of this research is to develop a technique for extracting the opinions from the online user reviews.

REVIEW OF RECENT RESEARCHES

Lots of researchers presented their views regarding opining mining. Few of views are described below.

Vu *et al*. (2011) have worked on opinion mining and hence introduced a Feature-based Opinion Mining and Summarizing (FOMS) of reviews. It considered reviews on mobile phones in Vietnamese to develop an opinion mining model. Similar to synonym features, direct/indirect feature-words and opinion-words have been retrieved and grouped into feature and stored in feature dictionary. VietSentiWordNet have been used to determine from features based on opinion orientation and summarization of customers using dedicated mathematical formulations.

Wonga and Lam (2009) have proposed an approach based on undirected graph model in which the mutual relationships between fragmental texts that belong to similar and different web pages have been modeled. Knowledge from various sources has been extracted using parallel consideration of web pages from different web sites. Retrieving information and mining the features have been performed by carrying out inference over the graphical model using a developed approximate learning algorithm. Two applications, mining significant product feature from vendor sites and hot item feature from auction sites, have been used to prove the framework efficiency. Effectiveness has also been reported by carrying out wide experimental study on the real-time data.

Liu *et al*. (2012) had introduced a new approach to retrieve opinion targets on the basis of word based translation model (WTM). Firstly, WTM has been applied in a monolingual condition to mine the relationship between opinion targets and words. Subsequently, candidate opinion relevance had been predicted from the mined relationship and included with significance of the candidate and hence derived a global measure, through which opinion targets had been

retrieved using a graph-based algorithm. WTM was able to identify the opinion relations accurately, more specifically for long term association. Compared with conventional syntax-based methods, WTM was more robust against noise that led from parsing errors occur in informal contents of big web copra. The graph-based algorithm has helped on global extraction of opinion targets in order to mitigate the error propagation problem that usually occurs in conventional bootstrap-based methods like Double Propagation. The effectiveness and robustness of the approach against state-of-the-art methods have been demonstrated using experimental outcome obtained from three real-time datasets in various sizes and linguistics.

Zhang and Liu (2011) have concentrated on extent verbs and adjectives, while less number of works had been carried out in the literature that deal with nouns and their phrases. When they have worked on feature-based opinion mining model for their research, they have discovered that in certain domains, opinions have also been represented by nouns and noun phrases that represent product features. These nouns remain objective rather than subjective in most of the cases. The sentences that include such nouns are also objective type and represent positive or negative opinions. In order to establish effective opinion mining in such domains, it is important to discover such nouns and their phrases and polarities, despite they are complex. The authors' method had dealt with this problem and it have outperformed when dealing with real-life datasets.

Kamal and Abulaish (2013) have introduced sentiment analysis to determine feature opinion pairs and their polarity by unifying rule-based and machine learning approaches. Experimenting with customer reviews on various electronic products had revealed the efficiency of the proposed method.

Zhai*et al*. (2011) have enhanced LDA, which is a renowned topic modeling method, to handle large scale constraints. Subsequently, they have introduced two new methods to automatically retrieve two varieties of constraints. Eventually, all the constraints that are obtained from the proposed and enhanced LDA have been combined to organize the product features. Constrained-LDA have been observed from the experimental results for its outstanding performance at large margin over conventional LDA and recent mLSA.

PROBLEM DEFINITION

Rapid growth on web technologies has led the users to generate huge volume of information in online systems, in the recent past. This huge volume of information has practical usage as data resources on applications related with sentiment analysis and opinion mining (Smeureanu and Bucur, 2012). The volume of information available in online systems has shown exponential increase with respect to the growth of internet and web 2.0 technologies that are driven by cost-effective technological infrastructure. Such huge volume of information is tedious to process individually. This resulted in information overload and often a bottleneck for decision-making in organizational processes. Hence, new techniques are required to generate knowledge for an organizational tactics (Wang and Wang, 2008).

Opinion mining is a subarea of data mining that plays a crucial role in discovering knowledge or retrieving information. The research interest on extracting the opinions of humans from web documents in an automatic manner has increased significantly. The primary intention of sentiment analysis is to help online customers on taking decision while purchasing any brand-new products. Opinion mining is a process of searching the sentiments of various users who have done online reviews, feedback, personal blogs, etc. The usage of blogs, reviews, etc has significantly increased because of today's wide utilization of Internet. Most of the users of such blogs, reviews are definitely customers of various online products. Due to wide global purchasing and sales of online products, the company has a serious concern on keeping their product up to date. Hence, the companies collect customer reviews on their products and decide on strength and weakness of the product using sentiment analysis (Jain *et al*., 2012).

Opinion mining can be facilitated by affluent data resource, which has been generated on web due to rapid growth on user-generated content. Nevertheless, effective retrieval of opinion still poses challenges to both individuals and entities because of reviews and their diversification on contents. Hence, automatic mining of opinions and summarization find crucial use (Miao *et al*., 2010). We use opinion mining for feature extraction by considering it as a potential tool to extract the reviews of the product from various web pages. Moreover, ratings based on the reviews not only assist the users to know about the product, but also it aid in comparing the product of different brands. Such feature extraction based on opinion mining consists of two major phases, namely, feature identification, which extracts and determines feature from user reviews and rating prediction phase that predicts L to define numerical rank for the product feature.

PROPOSED METHODOLOGY

The primary intention of this research is to develop a technique for extracting the opinions from online user reviews. The proposed methodology comprises of 4 major phases, which are:

- Data Pre-processing
- Potential feature extraction

- Opinion extraction and mining
- Opinion classification

Initially, the data, which are extracted from the web document, remain unstructured. The initial phase formats the data before performing sentiment analysis and mining. In the proposed work for feature extraction of the product, opinion mining is used to identify and compare the strengths and the weaknesses of the products based on the feedback given by the users on user reviews. Feature extraction is a crucial step for opinion mining, which has been used to collect the useful information from user reviews. The system that takes these features as input assigns ranks to them and decides final classification of the review as positive, neutral, or negative. The proposed algorithm to identify the features is called the Improved High Adjective Count (IHAC). This is done by Artificial Bee Colony (ABC) optimization algorithm. The main idea behind the algorithm is concentrating on the nouns for which reviewers express a lot of opinions and distinguishing them from nouns for which users do not express such opinions. After processing all reviews, this algorithm gives score for each noun. The ranking is used to find the scores that are greater than a threshold. The second proposed algorithm is Max opinion score algorithm, which ranks the extracted features using opinion scores assigned from previous method.
This algorithm comprises of three inputs such as:

- The list of adjectives, which are used to express opinions
- A score, which indicates the degree of positivity or negativity of the opinion
- The list of potential features

These can be identified using algorithms like the proposed IHAC. Finally, an evaluation is used for the proposed algorithms for their feature extraction from different review sites. The proposed framework not only classifies a review as positive or negative, but also extracts the most representative features of each reviewed item and assigns opinion scores for them. The final phase will be done using decision tree classifier with the help of extracted features.

Data pre-processing: Primarily, the dataset is collected from various web documents, which are the inputs for the proposed method. The unstructured or inefficiently structured data sources are retrieved and utilized further for data storage or data processing.

Pre-processing: The unstructured or inefficient raw input is appropriately subjected to the pre-processing method and the resultant is utilized for further processes. The pre-processing is mainly carried out with three steps that include:

- Stop words removal
- Stemming process
- POS representation

After the completion of these processes, the set of considerable Parts of Speech words are obtained as an output from every document by removing the stop words and stemmed processing of the words.

Stop words removal: This is the foremost step in the Pre-processing stage in which the stop words removal step filters out only the required words against a Stop Word list in order to eliminate the words, which are deemed to be unimportant for the user's opinion and feature of a product. The major reason for removing stop words is to preserve the system resources by eliminating those terms that contain small value for mining performance. The common words that are mentioned as stop words contain the function word sets and further more (i.e., articles, conjunctions, interjections, prepositions, pronouns and 'to be' verb forms). Usually, around 400-500 stop words are used in English language. The stop words like a, an, is, was and, the, to etc., are removed. After removing the human faults, the unnecessary words are removed.

Stemming process: The method of obtaining root words from the resultant words from the stop word removal process is known as stemming. In other words, Stemming is the task of removing the morphological and in-flexional ending words. The words are transformed into their stems by the Stemming process. The hypothesis, behind stemming, is that words with alike stem or word root will explain same or relatively close concepts in text and therefore words may be blended by means of stems. For instance, the words like user, users, used, using all may be stemmed to the word 'USE'.

POS Representation: Each of the individual sentences is represented using a WordNet, which assigns Parts-Of-Speech (POS) tags to the words based on the context in which each of the words appears.

WordNet: Wordnet is an English lexical database, which was constructed under the supervision of Miller (1995) at Princeton University and made available online. In the database, synsets such as nouns, verbs, adjectives and adverbs are grouped together to form the group of cognitive synonyms, where each synset represent diverse concepts. The synsets have been defined based on the conceptual, lexical and semantic relations. Wordnet can also be seen as ontology, when dealing with natural language terms. It contains 100000 words bystructuring them as taxonomic hierarchies.

Synonym sets (synsets) are the categorization of nouns, verbs, adjectives and adverbs. They are classes of senses, i.e., similar words or concepts with diverse synonyms. They have mutual connection at higher or lower degree based on the relationships between those synsets. Generally, the relationships can be said as two. They are Hyponym/Hypernym (i.e., Is-A relationships) and Meronym/Holonym (i.e., Part-Of relationships). Hyponym and Hypernym are referred as secondary organizing principle. If a word is said as the hyponym of another word, then the earlier world has more specific definition than the latter, whereas the latter can be said as the hypernym of the earlier. They exhibits transitive and opposite relationship each other. Hyponym/Hypernym is comprised of nine nouns and many verbs, apart from adjectives and adverbs (Varelas*et al*., 2005). When latest versions of ontology like Open Directory are released, the synset identifications may be modified. However, mapping of synsets among those versions can be done by using a backward compatibility utility program (Reed and Lenat, 2002).

Moreover, the file of reviews that corresponds to a particular product is divided into text files with one review in each file. Thus, the pre-processing stage provides the output as part-of-speech tagged text for a given input data.

Potential feature extraction: The part-of-speech tagged text is then processed to extract the potential features for which an Improved High Adjective Count (IHAC) Algorithm is utilized. The Improved High Adjective Count Algorithm is comprised of two algorithms, which are:

- High Adjective Count (HAC) Algorithm
- Artificial Bee Colony (ABC) algorithm

High adjective count algorithm: The algorithm, which identifies potential features, is known as High Adjective Count algorithm (HAC). This HAC algorithm helps to extract the most important features by which the opinions of reviewers can be expressed. The Nouns are the main part-of-speech tag that can express the features of a particular product. In general, most of the researchers utilized term-frequency of keywords. Instead of term-frequency, our proposed method uses Nouns and Adjectives for which we employ POS tag of a review document.

Finding opinion scores: Adjectives are the main concern in this High Adjective Count Algorithm and the counts of every adjective further helps to identify the potential features.

Each review is processed one by one to find the opinion scores. For any one of the review, the step-wise process is carried out. When a process is started, the adjectives and nouns are found out and the scores of all the nouns and the counts of all the adjectives are initially set to zero. The score of the noun is increased by one, if any adjective is close to the particular noun. Thus, each noun in all the reviews of the collected input database obtains scores, which are referred to as opinion scores.

Finding potential features: The nouns, which have more number of adjectives, are called as higher ranked nouns. The ranking process of noun is done by the opinion scores of each noun. This ranking process filters out the nouns and by which the potential features are chosen from the collection of nouns. In order to extract the potential features, a threshold is set for this algorithm. The nouns, which have scores greater than a threshold, are extracted as potential features.

Thus, the potential features can be extracted by using High Adjective Count algorithm with the aid of nouns and adjectives, rather than term-frequency of keywords.

Improved high adjective count algorithm: In the proposed work, we utilize Improved High Adjective Count Algorithm, which effectively optimize the scores of the nouns to extract the potential features. After finding opinion scores as mentioned in above section, we should optimize the scores and then rank every noun using Artificial Bee Colony Algorithm.

Artificial bee colony algorithm: ABC algorithm is a swarm based meta-heuristic algorithm, which is enthused by the sharp foraging behavior of the honey bees. It consists of three components namely, employed bees, onlooker bees and scout bees.

The employed bees are coupled with food sources in the nearby region of the hive and they transfer the data about the nectar quality of the exploiting food sources to the onlooker bees. Onlooker bees are watching the dance of the employed bees inside the hive to pick one food source to exploit based on the data provided by the employed bees. The employed bees, whose food sources are abandoned, become Scout bees and seek new food sources arbitrarily. The number of food sources denotes the location of probable solutions of optimization problem and the nectar amount of a food source denotes the quality of the solution. The working procedure of ABC algorithm is shown in Fig. 1.

Index value of nouns and its corresponding rank values are given as inputs to this Artificial Bee colony algorithm.

Initial phase: First, a population of the food sources x_i, (i = 1, 2,….R) is generated arbitrarily. R denotes the size of thepopulation. This foodsources contain the

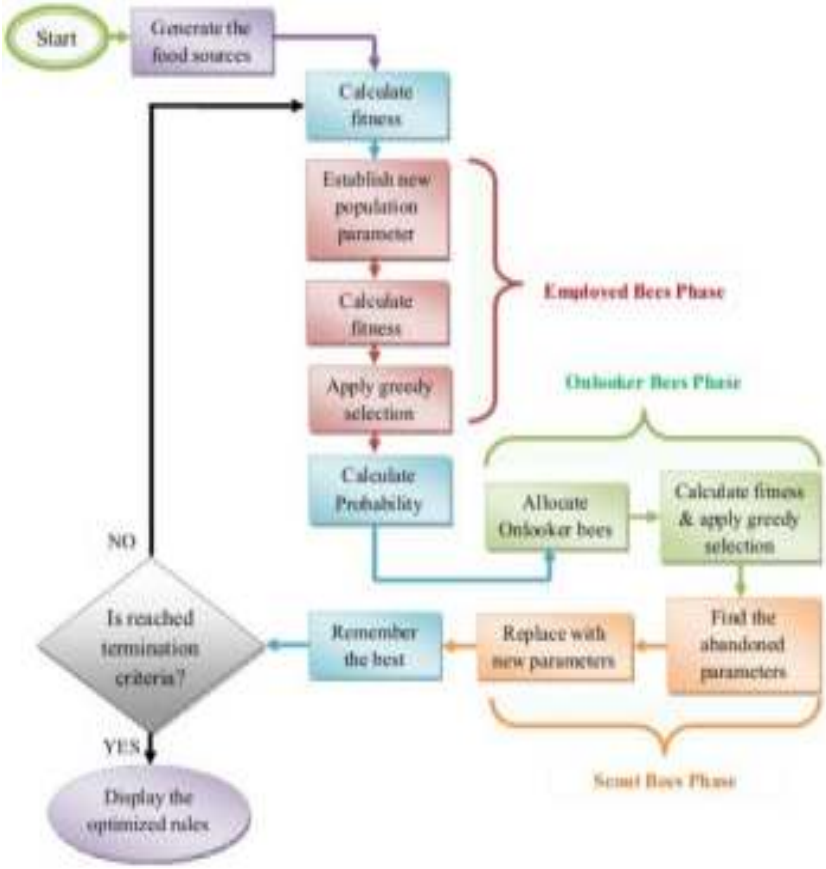

Fig. 1: Flowchart for ABC algorithm

index and corresponding rank values (R_i) that are generated for each noun. This generation process is called initialization process. To evaluate the best food source, the fitness value of the generated food source is calculated using Eq. (1):

$$Fitness\ Function, F(j) = \max(Rank) \qquad (1)$$

After the calculation of fitness value, the iteration is set to 1. Then, the phase of employed bee is carried out.

Employed bee phase: In the employed bee phase, new population parameters are generated using the equation below:

$$V_{i,j} = x_{i,j} + \phi_{ij}\left(x_{i,j} - x_{k,j}\right) \qquad (2)$$

where, *k*and*j* are randomly selected indices, ϕ is a randomly produced number in the range [-1, 1] and $V_{i,j}$ is the new value of the j^{th} position. Then the fitness value is computed for every newly generated population parameter of food sources. From the computed fitness values of the population, best population parameter is selected, i.e., the population parameter, which has the highest fitness value by applying greedy selection process. After selecting the best population parameter, probability of the selected parameter is computed using Eq. (3):

$$P_j = \frac{F_j}{\sum_{j=1}^{d} F_j} \qquad (3)$$

where,$P_j$is the probability of the j^{th}parameter.

Onlooker bee phase: After computing the probability of the selected parameter, number of onlooker bees are estimated. Later, new solutions $V_{i,j}$for the onlooker bees are generated from the solutions $x_{i,j}$based on the probability value P_j. Then, the fitness function is calculated for the new solution. Subsequently, the greedy selection process is applied to select the best parameter.

Scout bee phase: In this phase, the abandoned parameters for the scout bees are determined. If any abandoned parameters are present, then they are replaced by new parameters discovered by scouts using Eq. (3) and the fitness values are evaluated. The achieved best parameters so far are memorized. Then

the iteration is incremented and the process is continued until the stopping criterion is met. Finally, the optimized ranks and their corresponding indices are discovered.

Thus, the optimized ranks are obtained from the output of ABC algorithm and are further utilized to extract the potential features by checking the score values with threshold. The processes in above section are then followed by this ABC algorithm for finding the potential features. Hence, we can attain required potential features using Improved High Adjective Count Algorithm.

Opinion extraction and mining: After identifying the potential features, the opinion words are extracted and the mining process is performed using Max Opinion Score Algorithm. The three ways to extract and mine opinions are given below:

- Indentifying opinion words and assigning scores
- Concern about inversion words
- Potential features-opinion words matching

Indentifying opinion words and assigning scores: The adjectives, which are closest to the noun and extracted potential features are given as inputs to this step. These groups of adjectives, which are used to express opinions, are called as opinion words. Then, the scores are assigned for all the adjectives in the groups by which the nature of the opinion can be identified as positive or negative. Based on the opinion word, the scores are assigned. The negative scores show that the opinion is negative opinion words and vice versa. Positive opinion words have the higher values, whereas, negative opinion words have lower values.

Concern about inversion words: Inversion words can change the sense of the opinion word, which leads us to consider inversion words in the collection of opinion words. Inversion words are “not”, “un”, “non”, etc.,

For example, “good” indicates positive opinion word. However, if the inversion word “not” is added before this positive opinion word, then it will become “not good”, which signifies negative opinion.

Owing to this reason, when we are giving scores to the opinion words, we also keep the left context. If any inversion words are found as opinion word, then the score of that opinion word is multiplied by -1. Thus, the scores are assigned for the inversion opinion words.

Potential features-opinion words matching: The extracted Potential features are the inputs for matching the opinion words, which are found nearest to these potential features. We extract the potential features by the Improved High Adjective Count Algorithm.

Each review is sentence-wise processed. The opinion words and the potential features that are closest to every opinion word are identified for each sentence. Then the scores for every potential feature are gained by adding all the scores of opinion words that are related with particular potential feature. Subsequently, all these obtained scores of the whole potential features are summed up to compute the score for a review. Finally, the average score per opinion word is found for every potential feature.

The whole process of phase opinion extraction and mining using Max Score Algorithm is given as a pseudo code in the following:

Pseudo code for max opinion algorithm
Inputs: Reviews
Potential features
Adjectives
Adjective counts
Opinion scores
Outputs: Average scores
Assumptions: R_N→Review Noun scores
R_{NA}→Review Noun Adjective counts
LC→Left Context
LS→Line Scores
Pseudocode:
Begin
For each Review
Initialize $R_N = 0$ *and* $R_{NA} = 0$;
For each Line in a Review
Initialize $LC = 0$ *and* $LS = 0$;
For each Word in Line
If Word in Adjective scores
Then Score = Adjective score [Word];
If Inversion Words in LC
*Then Score = -1 * Score;*
Check for Closest Noun[Word]
$R_N[Closest\ \ Noun] += Score$;
$R_{NA}[Closest\ \ Noun]++$;
$LS += Score$;
Update LC ;
End
End

$$Total\ Score = \sum R_N \ ;\ Total\ \ Adjectives\ \ = \sum R_{NA}\ ;$$

$$Average\ \ Score = \frac{Total\ Score}{Total\ Adjectives}\ ;$$

End
Stop

Opinion classification: This phase is helpful for classifying the reviews, which have three kinds of opinions about a product. The classification is done by one of the classification algorithms named ID3 Algorithm. Based on the classification of opinions, we can separate the reviews into positive, negative and neutral. These three classes are related with the opinions.

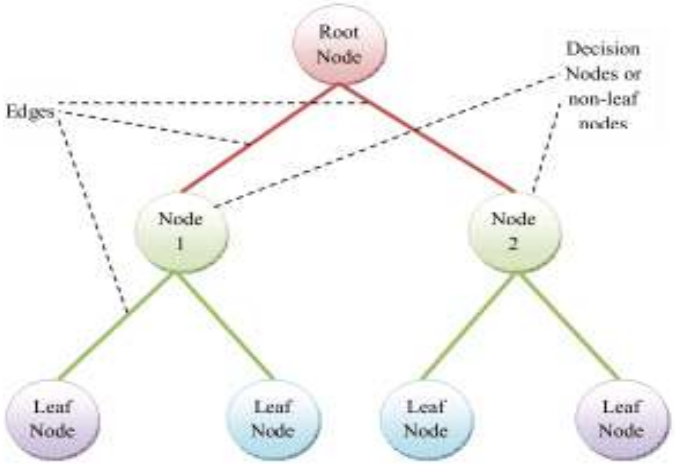

Fig. 2: General structure of decision tree diagram

ID3 Algorithm: ID3 (Interactive Dichotomizer version 3) is a simple decision tree learning algorithm that is used for classification purpose. It builds decision trees based on greedy search in an up-down manner, which contains the set of nodes and edges that connect these nodes. The leaf nodes of the decision tree have class names and the non-leaf nodes are the decision nodes, which take decision about its class labels. Every non-leaf node corresponds to an input attribute and every edge refers a possible value of that attribute. The decision nodes or the non-leaf nodes made an attribute test on every attribute of every tree node being a possible value of the attribute, through the given input sets. If the attribute classifies the given input sets perfectly, then the ID3 stops its process. Otherwise, it recursively operates on all the attributes in the decision tree to get the "best" attribute. The decision node to which the attribute goes from the non-leaf node is decided by the information gain metric of ID3. The resultant tree provides the classification of the samples given as inputs to this tree. The general structure of the decision tree diagram is given in Fig. 2.

According to our work, the result of classification should be any of the three- (1) Reviews that are Positive (set of positive results-PR) (2) Reviews that are Negative (set of negative results-NR) and (3) Reviews that are Neutral (set of neutral results-NTR). The total classes are represented as $T = PR \cup NR \cup NTR$. The number of sets of positive results (PR) is denoted as *a*, the number of sets of negative results (NR) is denoted as *b* and the number of sets of neutral results (NTR) is denoted as c.

The intermediate information required to make the decision tree as a source of the result PR or NR or NTR is given as:

$$I(a,b)=\begin{cases} -\frac{a}{a+b}\log_2\frac{a}{a+b} - \frac{b}{a+b}\log_2\frac{b}{a+b}, & when\ \ a,b\neq 0 \\ 0 \quad , & otherwise \end{cases} \quad (4)$$

$$I(b,c)=\begin{cases} -\frac{b}{b+c}\log_2\frac{b}{b+c} - \frac{c}{b+c}\log_2\frac{c}{b+c}, & when\ \ b,c\neq 0 \\ 0 \quad , & otherwise \end{cases} \quad (5)$$

$$I(a,c)=\begin{cases} -\frac{a}{a+c}\log_2\frac{a}{a+c} - \frac{c}{a+c}\log_2\frac{c}{a+c}, & when\ \ a,c\neq 0 \\ 0 \quad , & otherwise \end{cases} \quad (6)$$

If attribute (feature) X_i with values $\{v_1, v_2, ...v_N\}$ is used for the root of the tree, then it partitions *T* into $\{T_1, T_2, ...T_N\}$, where T_i denotes the classes in *T* that have value v_i of X_i. Let T_i consists of a_i classes of PR; b_i classes of NR and c_i classes of NTR, then the intermediate information needed for the sub-tree for T_i are $I(a_i,b_i)$, $I(b_i,c_i)$ and $I(a_i,c_i)$. The expected information *EI* needed for the decision tree with X_i as the root ($EI(X_i)$), is found as a weighted average as given below.

$$EI(X_1) = \sum_{i=1}^{N} \frac{a_i + b_i}{a+b} I(a_i,b_i) \quad (7)$$

$$EI(X_2) = \sum_{i=1}^{N} \frac{b_i + c_i}{b+c} I(b_i,c_i) \quad (8)$$

$$EI(X_3) = \sum_{i=1}^{N} \frac{a_i + c_i}{a+c} I(a_i,c_i) \quad (9)$$

From the above Eq. (7)-(9), the weight for the i th branches is the proportion of the classes *T* that belongs to T_i. Now, we can find the value of information gain IG by branching on X_i, which is given below:

$$IG(X_1) = I(a,b) - EI(X_1) \quad (10)$$

$$IG(X_2) = I(b,c) - EI(X_2) \quad (11)$$

$$IG(X_3) = I(a,c) - EI(X_3) \quad (12)$$

All the features X_i are analyzed by ID3 algorithm and X_i, which gives maximum $IG(X_i)$ value is selected for constructing the decision tree. Then for the residual subsets T_1, T_2,…T_N, the same process recursively continues to construct the tree. For each T_i, (i=1, 2, …N), if all the classes are positive, then it makes a decision with "YES" node and stops the process and if all the classes are negative, then it makes decision with "NO" node and stops the process; Otherwise ID3 chooses another attribute in the same way as given earlier that makes the Neutral class with "Neutral" node.

Thus the reviews are classified into Positive, Negative and Neutral based on the opinions of the users that are included in the reviews.

RESULTS AND DISCUSSION

Our proposed Improved High Adjective Count algorithm based opinion mining method is implemented in Java platform. The working procedure for our proposed work is explained in the previous section and the results for the implementation of our work are given in this section.

Data sets description: Two datasets are taken for our proposed work and in which same product with different kinds are chosen for our work. The two datasets are:

- Customer review datasets
- Additional review datasets

Customer review datasets: This dataset contains totally reviews of 5 products. The products are:

- Apex AD2600 Progressive-scan
- Canon G3
- Creative Labs Nomad Jukebox
- Nikon CoolPix 4300
- Nokia 6610. From these 5 products, we choose only two camera models Canon G3 and Nikon CoolPix 4300.

Additional review datasets: In this dataset, the reviews about 9 products are enclosed:

- Canon PowerShot SD500
- Canon S100
- Diaper Champ
- Hitachi router
- ipod
- Linksys Router
- MicroMP3
- Nokia 6600
- Norton are the products

Which are comprised in this dataset. Among these 9 product reviews, only 2 reviews are selected for our proposed work. The datasets used for our work are Canon PowerShot SD500 and Canon S100.

[t]excellent picture quality / color
canon power shot g3[+3]##i recently purchased the canon power shot g3 and am extremely satisfied with the purchase.

use[+2]##the camera is very easy to use, in fact on a recent trip this past week i was asked to take a picture of a vacationing elderly group.
##after i took their picture with their camera, they offered to take a picture of us.
##i just told them, press halfway, wait for the box to turn green and press the rest of the way.

picture [+2]##they fired away and the picture turned out quite nicely. (As all of my pictures have thusfar).

picture quality[+1]##a few of my work constituents owned the g2 and highly recommended the canon for picture quality.

picture quality[+1]##i 'm easily enlarging pictures to 8 1/2 x 11 with no visible loss in picture quality and not even using the best possible setting as yet (super fine).
##ensure you get a larger flash, 128 or 256, some are selling with the larger flash, 32mb will do in a pinch but you'll quickly want a larger flash card as with any of the 4mp cameras.

camera[+2], use[+2], feature[+1]##bottom line, well made camera, easy to use, very flexible and powerful features to include the ability to use external flash and lense / filters choices.

picture quality[+3], use[+1], option[+1]##i 'd highly recommend this camera for anyone who is looking for excellent quality pictures and a combination of ease of use and the flexibility to get advanced with many options to adjust if you like.
##great job canon

Fig. 3: Sample dataset-Canon G3 of customer review dataset

All the reviews of the products were from amazon.com. These two datasets are collected from this link: http://www.cs.uic.edu/~liub/FBS/senti ment-analysis.html

Experimental results: In this section, the results for our proposed implementation is shown and explained in detail with experimentation results. Initially, the dataset sample is given in Fig. 3 for the Canon G3 of Customer Review Dataset.

Preprocessing is the first phase for our work, the results of preprocessing stage for the dataset in Fig. 3 is given in Table 1.

Followed by the pre-processing phase, our proposed Improved High Adjective Count Algorithm employs on the Noun words, which are considered as the features of opinion mining work. The Nouns and its score are given in the following Table 2.

The sample results of third phase, opinion mining and extraction are tabulated in Table 3 for the whole Canon G3 data. Nouns from IHAC algorithm are considered as features and its relevant opinion words are obtained from the adjectives.

Based on these opinion words, we classify the whole review datasets and can identify the particular product is good or bad by the review classification results. The whole datasets are processed by our proposed method and the final sample of review classification. Here, the results of four camera types Canon G3, Canon S100, Nikon coolpix 4300 and Canon Powershot SD500 are given in Table 4.

Evaluation metrics: An evaluation metric is used to evaluate the effectiveness of opinion mining systems and to justify theoretical and practical developments of these systems. It consists of a set of measures that follow a common underlying evaluation methodology. Some of the metrics that we have chosen for our evaluation purpose are Recall, Precision and the F-measure.

In order to employ our proposed method for the effective classification of reviews with mining, we require these evaluation metric values to be computed. The metric values are found based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) with the possibility of classification of reviews. Table 5 shows how the positive and negative values are described.

Using these four basic values, the metrics of Precision, Recall, F-Measure and Accuracy are calculated in our proposed method. The representation of these evaluation metrics are given below in equations.

Precision: The precision estimates how many of the reviews classified to be Positive (Negative or Neutral) are actually Positive (Negative or Neutral) by means of the equation:

$$\text{Precision} = \frac{TP}{FP + TP} \tag{13}$$

Table 1: Preprocessing phase results

Stemming	[Satisfy, satisfied]		[Recommend, recommended]	
	[vacation, vacationing]		[set, setting]	
	[fire, fired]		[sell, selling]	
	[turn, turned]		[make, made]	
	[own, owned]		[advance, advanced]	
POS Representation	[recently, adverb]	[halfway, adjective]	[quality, adjective]	[features, noun]
	[purchased, verb]	[wait, noun]	[setting, noun]	[ability, noun]
	[canon, noun]	[box, noun]	[super, adjective]	[external, adjective]
	[extremely, adverb]	[turn, noun]	[fine, adjective]	[flash, adjective]
	[satisfied, adjective]	[press, noun]	[larger, adjective]	[lense, noun]
	[purchase, noun]	[rest, noun]	[flash, adjective]	[filters, noun]
	[camera, noun]	[fired, adjective]	[selling, noun]	[choices, noun]
	[easy, adjective]	[picture, noun]	[larger, adjective]	[highly, adverb]
	[fact, noun]	[turned, adjective]	[flash, adjective]	[recommend, verb]
	[recent, adjective]	[nicely, adverb]	[mb, noun]	[camera, noun]
	[trip, noun]	[picture, noun]	[pinch, noun]	[excellent, adjective]
	[past, adjective]	[work, noun]	[quickly, adverb]	[quality, adjective]
	[week, noun]	[owned, adjective]	[larger, adjective]	[picture, noun]
	[asked, verb]	[highly, adverb]	[flash, adjective]	[combination, noun]
	[picture, noun]	[recommended, adjective]	[card, noun]	[ease, noun]
	[vacationing, noun]		[mp, noun]	[flexibility, noun]
	[elderly, adjective]	[canon, noun]	[cameras, noun]	[advanced, adjective]
	[group, noun]	[picture, noun]	[bottom, adjective]	[options, noun]
	[picture, noun]	[quality, adjective]	[line, noun]	[adjust, verb]
	[camera, noun]	[easily, adverb]	[made, adjective]	[great, verb]
	[offered, verb]	[enlarging, verb]	[camera, noun]	[job, noun]
	[picture, noun]	[picture, verb]	[easy, adjective]	[canon, noun]
	[told, verb]	[loss, noun]	[flexible, adjective]	
	[press, noun]	[picture, noun]	[powerful, adjective]	

Table 2: Features and its scores by IHAC-sample results

Features	Scores
Canon	16
Purchase	2
Camera	67
Fact	1
Trip	1
Week	0
Picture	15
Vacationing	1
Group	0
Press	1
Wait	1
Box	3
Turn	1
Rest	1
Pictures	17
Work	3
Loss	0
Setting	2
Selling	2
Mp	0
Pinch	0
Card	5
Mp	3
Cameras	13
Line	2
Features	12
Ability	2
Lense	0
Filters	1
Choices	0
Combination	0
Ease	1
Flexibility	1
Options	3
Job	0

Recall: The recall indicates how many of the reviews of Positives (Negatives or Neutrals) classes actually are classified. The percentage of Positives (Negatives or Neutrals) correctly classified is represented using recall. It is also equal to Sensitivity:

$$\mathrm{Re}\,call = \frac{TP}{FN + TP} \tag{14}$$

Table 3: Opinion mining and extraction phase results

Features	Opinion word	Rank
Pictures	Mixed-light	17
Features	External	12
Money	Developing	1
Battery	Charged	4
Price	Recent	1
Lcd	Bottom	4
Flexibility	Advanced	1
Zoom	Hot	8
Screen	Great	2
Focus	Flash	2
Resolution	Worried	2
Control	Imaginable	1
Lens	Required	7
Buttons	Easy	1
Viewfinder	Stunning	3
Batteries	Proprietary	1
Setting	Maximum	2
Photo	Great	5
Shutter	Camera-determined	4
Display	Waste	1
Panel	Positive	2
Power	Advanced	1
Picture	Quality	15
Zooms	Closer	1
Photos	Highest	7

F-Measure: F-Measure combines precision and recall is the harmonic mean of precision and recall:

$$F - Measure = \frac{2\left(\mathrm{Pr}\,ecision \times \mathrm{Re}\,call\right)}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{15}$$

Accuracy: The weighted percentage of Positive, Negative and Neutral reviews that are correctly classified is measured by accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \tag{16}$$

Performance analysis: The evaluation results for our proposed Improved High Adjective Count based opinion mining work is illustrated in Table 6.

Table 4: Classification results of reviews

Description	Canon G3	Canon S100	Nikon cool pix 4300	Canon power shot SD500
Positive	25.00	49.00	25.00	1.0
Neutral	20.00	2.00	9.00	0.0
Negative	0.00	0.00	0.00	0.0
Result	Positive	Positive	Positive	Positive

Table 5: Description of TP, TN, FP and FN values

		OUTPUT	
Descriptions		Classified as positive	Classified as not positive
INPUT	Actually positive	TP	FN
	Actually not positive	FP	TN

Table 6: Performance evaluation of our proposed work

	Evaluation Metrics (in %)			
Product names	Precision	Recall	F-Measure	Accuracy
Canon G3	94.56	76.35	76.32	93.67
Canon S100	93.26	75.54	85.14	91.73
Nikon CoolPix 4300	95.35	86.32	84.26	94.24
Canon PowerShot SD500	94.63	84.52	84.35	92.56

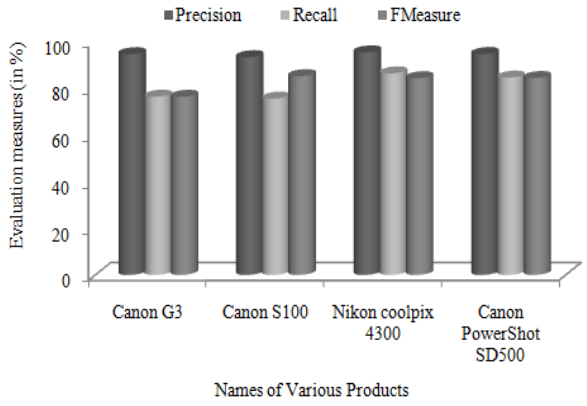

Fig. 4: Evaluation measures precision, recall and f-measure of various products for our proposed work

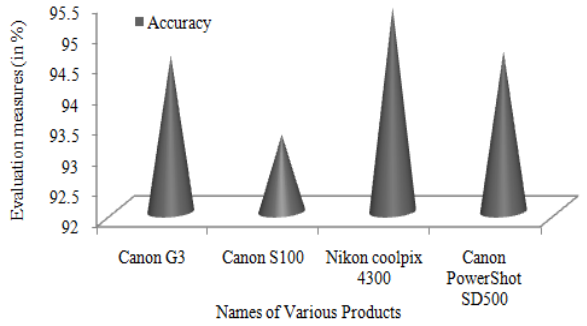

Fig. 5: Accuracy measures of various products for our proposed work

The performance analysis with the evaluation metrics Precision, Recall and F-measure are given in Fig. 4.

The reviews for the products of various companies are collected in our work with the aid of our proposed IHAC method. Here, we take only one product with various kinds of that particular thing. The products that are reviewed are Canon G3, Canon S100, Nikon CoolPix 4300 and Canon PowerShot SD500. These are the various kinds of the product camera, by which, we can find the review performance of our proposed work. For these four kinds of camera, the evaluation metrics precision, recall and F-Measure values are obtained by the description of TP, TN, FP and FN values. We can analyze the performance of our proposed work based on these evaluation metrics. Among the four kinds, Nikon CoolPIx 4300 model obtains very high precision value with 95.35% and the models Canon PowerShot SD 500, Canon G3 and Canon S100 have 94.63, 94.56 and 93.26% of precision, respectively. The recall values for these various kinds, Canon G3, Canon S100, Nikon CoolPix 4300 and Canon PowerShot SD500 are 76.35, 75.54, 86.32 and 84.52%, respectively. Nikon CoolPix 4300 model acquires the higher value (86.32%) for the metric recall. And also, Canon G3 attains lower F-measure value in compared with other three cameras.

Table 7: Comparison between proposed and existing opinion mining and classification methods

Methods	Precision
J48 (Kamal and Abulaish, 2013)	88.5
Bagging (Kamal and Abulaish, 2013)	87.5
Proposed work	94.45

Canon S100, Nikon CoolPix 4300 and Canon PowerShot SD500 gains 85.14, 84.26 and 84.35%, respectively of F-Measure values, in which Canon S100 get higher F-Measure value. On average, we achieve 94.45% of precision, 80.6825% of recall and 82.5175% of f-measures for our proposed work. This is very good value for the opinion mining of online reviews, which shows that our proposed Improved High Adjective Count algorithm effectively mines and classifies the reviews.

And also, the accuracy measure is given in Fig. 5 with the graphical representation for our proposed methodology.

With the assist of our proposed Improved High Adjective Count method, the assessments for the same product of various companies are brought together in our work. The goods that are reviewed for examining the measure accuracy are Canon G3, Canon S100, Nikon CoolPix 4300 and Canon PowerShot SD500. Accuracy measure values are also acquired by the description of TP, TN, FP and FN values. Comparing with all these types of camera products, Nikon CoolPIx 4300 model acquires higher accuracy value with 94.24% and the products Canon G3, Canon PowerShot SD 500 and Canon S100 have 93.67, 92.56 and 91.73% of accuracy, respectively. The average value for the accuracy is 93.05% of accuracy on average value for our proposed work. Thus, we achieve incredibly excellent accuracy values for the opinion mining and classification of reviews by the results and we can prove that our proposed Improved High Adjective Count algorithm successfully mines the opinions and classifies the online reviews.

Comparative analysis: Our proposed Improved High Adjective Count based Opinion Mining and classification method is compared with the existing methods ofopinion mining and classification (Kamaland Abulaish, 2013) in the literature survey of in above section. The comparison table for the proposed and existing opinion mining methods is given in Table 7.

The precision value for the proposed work is high for the mining and classification of opinions. IHAC and ID3 algorithms together yields 94.45% of precision on average for the proposed work, which shows that the usage of both IHAC and ID3 in the proposed work leads to make improvement in the opinion mining and classification process. The another two algorithms J48 and Bagging of state-of-art works facilitates only 88.5%

and 87.5% of precision value, by which we can prove that the proposed IHAC based opinion mining work is better for mining and classifying the opinions.

CONCLUSION

The proposed effective opinion mining and classification algorithm was carried out on Customer Review Datasets and Additional Review Datasets with the aid of JAVA platform. The products based on these two datasets were Canon G3, Canon S100, Nikon CoolPix 4300 and Canon PowerShot SD500. These products were reviewed the various kinds of the product camera, by which, the performance of the proposed work was found. According to the TP, TN, FP and FN values, the reviews were analyzed using the evaluation metrics precision, recall, f-measure and accuracy values. On average, 94.45% of precision, 80.6825% of recall, 82.5175% of f-measure and 93.05% of accuracy values were obtained for the proposed work. This is very good value for the opinion mining of online reviews, which shows that the proposed Improved High Adjective Count algorithm effectively mines and classifies the reviews. The existing two algorithms J48 and Bagging gives only 88.5% and 87.5% of precision value, by which we can demonstrate that the proposed IHAC based opinion mining work is good for mining and classifying the opinions.

REFERENCES

Chang, C.H., M. Kayed, M.R. Girgis and K.F. Shaalan, 2006.A survey of web information extraction systems. IEEE T. Knowl. Data En., 18(10): 1411-1428.

Etzioni, O., M. Cafarella, D. Downey, A.M.Popescu, T. Shaked, S. Soderland, D.S. Weld and A. Yates, 2005. Unsupervised named-entity extraction from the web: An experimental study. Artif.Intell., 165(1): 91-134.

Hong, J.L., 2010. Deep web data extraction.Proceeding of the IEEE International Conference on Systems Man and Cybernetics (SMC, 2010). Istanbul, pp: 3420-3427.

Jain, A., S. Jain, P. Shukla and H. Bandiya, 2012.Towards automatic detection of sentiments in customer reviews. Int. J. Inform. Sci. Tech., 2(4): 103.

Kamal, A. and M. Abulaish, 2013. Statistical features identification for sentiment analysis using machine learning techniques. Proceeding of the International Symposium on Computational and Business Intelligence (ISCBI, 2013). New Delhi, pp: 178-181.

Liu, K., L. Xu and J. Zhao, 2012.Opinion target extraction using word-based translation model. Proceeding of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12), pp: 1346-1356.

Meena, A. and T.V. Prabhakar, 2007. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. Proceeding of the 29th European Conference on IR Research (ECIR'07). Rome, Italy, pp: 573-580.

Miao, Q., Q. Li and D. Zeng, 2010.Mining fine grained opinions by using probabilistic models and domain knowledge.Proceeding of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT, 2010). Toronto, ON, 1: 358-365.

Miller, G.A., 1995. WordNet: A lexical database for English. Commun. ACM, 38(11): 39-41.

Pang, B. and L. Lee, 2008.Opinion mining and sentiment analysis. Found. Trend. Inf. Retriev., 2(1-2): 1-135.

Reed, S.L. and D.B. Lenat, 2002.Mapping ontologies into cyc.Proceeding of the AAAI Conference 2002 Workshop on Ontologies for the Semantic Web. Edmonton, Canada, pp: 1-6.

Smeureanu, I. and C. Bucur, 2012. Applying supervised opinion mining techniques on online user reviews. Informatica Econ., 16(2): 81.

Varelas, G., E. Voutsakis, P. Raftopoulou, E.G.M. Petrakis and E.E. Milios, 2005.Semantic similarity methods in wordnet and their application to information retrieval on the web. Proceeding of the 7th Annual ACM International Workshop on Web Information and Data Management, pp: 10-16.

Vu, T.T., H.T. Pham, C.T. Luu and Q.T. Ha, 2011. A Feature-based Opinion Mining Model on Product Reviews in Vietnamese. In: Katarzyniak, R. *et al*. (Eds.), Semantic Methods for Knowledge Management and Communication. Studies in Computational Intelligence, Springer-Verlag, Berlin, Heidelberg, 381: 23-33.

Wang, H. and S. Wang, 2008. A knowledge management approach to data miningprocess for business intelligence. Ind. Manage. Data Syst., 108(5): 622-634.

Wonga, T.L. and W. Lam, 2009.An unsupervised method for joint information extraction and feature mining across different Web sites.Data Knowl. Eng., 68(1): 107-125.

Zhai, Y. and B. Liu, 2006. Structured data extraction from the web based on partial tree alignment. IEEE T. Knowl. Data En., 18(12): 1614-1627.

Zhai, Z., B. Liu, H. Xu and P. Jia, 2011.Constrained LDA for grouping product features in opinion mining. Proceeding of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'11), 1: 448-459.

Zhang, L. and B. Liu, 2011.Identifying noun product features that imply opinions. Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, 2: 575-580.