A Survey On Weighted Sentiment Analysis using Artificial Bee Colony Algorithm

Ms. Ruby Dhurve, Asst Prof. Megha Seth

Abstract— **Data mining is the process of analyzing interesting data from different perspective and summarizing into useful information. Sentiment analysis is an application of natural language processing, data mining and text mining to identify sentiments or mood of the public about particular topic or products or customer reviews. This paper proposes to improve the methods for feature selection for customers review and also detect the polarity of reviews using machine learning approach then consider score as evidence for overall reviews weighting. Objective of the research paper is to select the best features selection methods for the aspect level of the sentiment analysis. Bog of noun, bog of words, stop word removal , parts of speech, ABC algorithm are used for feature set selection. For classification K-NN, Naïve Bayes, Support vector machine clustering algorithms are used for classification of sentiment weighted analysis.**

Index Terms— **Sentiment analysis; Feature selection; Artifiacal bee colony algorithm; Term weighting; Support Vector Machine.**

I. INTRODUCTION

Data mining is the process of analyzing interesting data from perspective and summarizing into useful information. Sentiment analysis or opinion mining is an application of natural language processing and text analysis to identify and extract sentiments from a give source. Sentiment Analysis used to identify the attitude, judgment, evaluation or emotional communication of a reviewer or speaker with respect to some topic in document [3]. Sentimental Analysis is all about to get the real voice of people towards specific product, services, organization, movies, news, events, issues and their attributes. Sentiment Analysis includes branches of computer science like Natural Language Processing, Machine Learning, Text Mining and Information Theory and Coding. Sentiment analysis is done on three levels Document Level, Sentence Level, Entity or Aspect Level. Document Level Sentiment analysis is performed for the whole document and then decide whether the document express positive or negative sentiment [7],[11].

The paper start on discussion in Section II Literature survey has been shown in section II. A discussion on ABC algorithm is presented in section IV. Finally, this papers ends with conclusion.

Manuscript received April, 2015.

Ms. Ruby Dhurve, *Computer Science & Engineering, RCET Bhilai, Bhilai India.*

Asst Rrof. Megha Seth*, Computer Science & Engineering, RCET Bhilai, Bhilai India.*

II. LITERATURE REVIEW

Earlier papers were describing the sentiment analysis or opinion mining on people's opinion expressed in written language or text and various clustering algorithms are illustrated and discussed below. [1] This paper describe aspect level sentiment analysis considering three classes for sentiment polarity of each sentence (positive, neutral and negative). In the aspect identification step they proposed to not ignore the part-of-speech tags, and instead of clustering with bag of words, employ a clustering over the sentences only using bag of nouns. Farhadloo.M et al (2013) results show that clustering with BON yields more meaningful aspects than using BOW. The main contribution of this paper is the proposal of a new feature set and score representation that leads to more accurate sentiment analysis by using SVM classifier. This scheme is based upon the three scores (positiveness, neutralness and negativeness) that are learned from the data for each term. Using this new score representation scheme, they improve the performance of 3-class sentiment analysis on sentences by 20% in terms of average f1-score, as compared to previously published research. This indicates that there is still room for feature engineering to improve the performance of classifiers in sentence-level opinion mining, and expect that future research will continue to improve on opinion mining[1]. The implications for practice from this paper are i) much improved performance of the sentiment analysis, and ii) an ability to more accurately extract sentiments from domains with higher granularity of opinions (positive, neutral, and negative).

The five different machine learning method and SVM sentiment classifier. Find J48 classification tree for highest performance, closely followed by Random Forest(RF), Support vector machine is used as the superior in the Domain. These exhaustive searches resulted in 5 features in the optimal feature subset when evaluated with the classifiers Naive Bayes (NB) and Random Forrest (RF), 6 features for Artificial Neural Networks (ANN) and J48 and 7 for Support Vector Machines (SVM).In that research found features relying on context specific sentiment lexica ('contextual' category) to be paramount in classification within this domain – yielding a precision increase of ~21% when added to the feature categories[2].

Valakunde.N et al (2013) performed hierarchical sentiment analysis. Document level sentiment is computed by calculating aspect level sentiment score and the

corresponding weightages given to the entities. When performed aspect based document level sentiment analysis accuracy is high as compared to sentiment analysis at direct document level. Instead of giving information in terms of only positive and negative class our approach gives a multiclass sentiment analysis providing fine-grained view of sentiments. The result shows that negation handling is very important as they leads to misclassify the documents. It also shows that SVM has better accuracy over NB.

Liu.L et al (2012) research A Fuzzy Domain Sentiment Ontology Tree can be automatically constructed to facilitate opinion mining, including the extraction of product features and sentiment words, extraction of feature-relation. That method can accurately predict the polarities of sentiments. As a result, organizations can develop effective business strategies related to marketing, customer support, and product design functions in a timely fashion.

T.Sumathi et al (2013) Optimal feature selection is used for reducing feature subset size and computational complexity thereby increasing the classification accuracy. The ABC algorithm being a powerful optimization technique and is widely used for solving combinatorial optimization problems. Hence, this method is incorporated for optimizing the feature subset selection in this investigation. In this paper, the movie reviews is classified using opinion mining. Experiment results evaluated feature selection techniques based on IDF and proposed ABC. Experimental results show that the classification accuracy of the classifiers improves in tune of 1.63% to 3.81% for the proposed ABC feature selection.

Basari.A et al (2013) has shown that PSO affect the accuracy of SVM after the hyridization of SVM-PSO. The best accuracy level that gives in this study is 77% and has been achieved by SVM-PSO after data cleansing. On the other hand, the accuracy level of SVM-PSO still can be improved using enhancements of SVM that might be using another combination or variation of SVM with other optimization method.

Manjul.S et al (2014) Probabilistic aspect algorithm is proposed using K-NN based sentiment classifier. In this proposed system, semantic-oriented subjective information is first extraction and raking is calculated based on the aspect frequency

Vigneshkumar.K et al (2013) learn so many classification techniques such as k-Nearest Neighbor Classifier Support Vector Machine, and Principal Component Analysis, Artificial neural network, Fuzzy logic. This paper provides an overview to improve the prediction accuracy to enhance the quality factor consider ARSQA, an Autoregressive sentiment and Quality Aware model to build the quality for predicting sales performance.

Basiri.M et al (2014) propose a new score aggregation method based on the Dempster-Shafer theory of evidence. In the proposed method, and first detect the polarity of reviews using a machine learning approach and then, consider sentence scores as evidence for the overall review rating. The results from two public social web datasets show the higher performance of that method in comparison with existing score aggregation methods and state-of-the-art machine learning approaches.

Tripathi.G etal (2014) presents a survey on sentiment analysis and the related techniques. It also discusses the application areas and challenges for sentiment analysis with insight into the past researches. S.Joshi.N et.al,(2014)This paper provides a study of the most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms. The NB classifier produces the best results, followed by the DT classifier.

III. Problem Identification

As the size of digital information grows exponentially, large volumes of raw data need to be extracted. Nowadays, there are several methods to customize and manipulate data according to our needs. The most common method is to use Data Mining (DM). DM has been used in previous years for extracting implicit, valid, and potentially useful knowledge from large volumes of raw data (Sousa, Silva, & Neves). The extracted knowledge must be accurate, readable, comprehensible, and ease of understanding. Furthermore, the process of data mining is also called as the process of knowledge discovery which has been used in most new inter-disciplinary area such as database, artificial intelligence statistics, visualization, parallel computing and other fields. We found that many optimization algorithms have been used for classification tasks. From the best of our knowledge, previous researches on ABC algorithm have focused on optimization but none of them is for classification tasks [19].

The ABC algorithm is innovated in 2005 by Karaboga inspiring the social life of the bees to solve the optimization problems. This algorithm is a simulation of the food search of the group of the bees [22]. The group bees can be distributed in different distances to utilize the food resources. In ABC algorithm, the bees are classified in three groups:

1. Employed bees,
2. Onlooker bees,
3. Scout bees.

In previous research paper have describe the various methods for selecting for feature set using bag of word, bag of noun, parts-of-speech, stemmer etc has very complex to select the feature set from sentiment reviews of customer. We have use ABC algorithm to optimize the feature selection of subset size. It reduce the complexity in selecting subset of feature that are extracted from the sentiment text or sentences.

IV. Methodology Used

A. Aspect Identification

In this section we will describe the method that we will use to extract the aspects of the object and the preprocessing steps involved in preparing the sentences for clustering. We use the term “review” in order to describe a user-generated comment The idea behind the use of clustering techniques, is to find the aspects of the object that users have expressed their opinions in the reviews. That is, if we have access to all sentences of all reviews and somehow are able to find similar sentences, it is likely that similar sentences are about similar aspects. If we can properly represent each sentence of the review with a vector, then if we apply the clustering algorithm to these feature vectors, the sentences in each cluster are similar

sentences that are probably addressing the same aspect of the object [1]. As a result they prposed clustering approach which finds the cluster by shorting the weighted term frequency of all the terms.

The major reason for failure of the regular clustering algorithms in the experiments, is that the lack of using a poor method to represent each sentence before applying clustering. Farhadloo.M et al (2013) use Bag of Noun instead of Bag of Word and Part Of Speech tag for similar sentence in clustering.

B. Sentiment Identification

In aspect level sentiment analysis, after identifying the aspect, the sentiment of each sentence containing one of those aspects has to be identified [1]. On that paper only two approach is used for to identify sentiment of sentence that are classifier and feature extractor.

For feature extraction use Bag of word with Artificial Bee Colony Algorithm to minimize the feature selection subset size and computational complexity thereby increasing the classification accuracy, ABC algorithm is used for feature reduction.

The artificial bee colony algorithm is inspired by the behavior of the bee colony in nectar collection. This biologically inspired approach is currently being employed to solve continuous optimization problems, training neural networks, mechanical and electronically components design optimization, combinatorial optimization problems such as job shop scheduling, the internet server optimization problem, etc.

There are three phase in ABC algorithm

1.The Employed Bee Phase

2.The Onlooker Bee Phase

3.The Scout Bee Phase

C. Algorithm for ABC Optimization

T.Sumathi et al (2014) proposed ABC algorithm is used to optimize the feature selection as follows:

1. Cycle =1
2. Initialize ABC parameters
3. Evaluate the fitness of each individual feature
4. Repeat

a. Construct solutions by the employed bees

i. Assign feature subset configurations to each employed bee

ii. Produce new feature subsets

iii. Pass the produced feature subset to the classifier

iv. Evaluate the fitness of the feature subset

v. Calculate the probability of feature subset solution

b. Construct solutions by the onlookers

i. Select a feature based on the probability

ii. Compute vi using xi and xj

iii. Apply greedy selection between vi and xi

c. Determine the scout bee and the abandoned solution

d. Calculate the best feature subset of the cycle

e. Memorize the best optimal feature subset

f. Cycle=Cycle+1

Until pre-determined number of cycles =1000 or Root Mean Square Error (RMSE)<0.01

5. Employ the same searching procedure of bees to generate the optimal feature subset configurations.

Each employed bee is moved to food source area to determine a new food source in the neighbourhood of the current one, and its nectar amount evaluated. If nectar amount of new source is higher, then the bee forgets the first source and memorizes the new one. Onlookers are placed on food sources by using probability based selection process. As nectar amount in food source increases, probability value with which it is preferred by onlooker's increases similar to natural selection process in evolutionary algorithms [5].

D. Weight Representation

In weight based representation three weight are computed for each term (t_i) in our vocabulary list: positive score (s_i^+), neutral score (s_i^0) and negative score (s_i^-) [1]. These scores are computed as:

$$S_i^+ = \frac{f_i^+}{(f_i^+ + f_i^0 + f_i^-)},$$

$$S_i^0 = \frac{f_i^0}{(f_i^+ + f_i^0 + f_i^-)},$$

$$S_i^- = \frac{f_i^-}{(f_i^+ + f_i^0 + f_i^-)}$$

where f_i^+, f_i^0, f_i^- are the frequencies of term t_i in positive, neutral and negative documents respectively. Using these scores, [1] compute the positiveness, neutralness and negativeness of each sentence (x) as:

$$s^+ = \sum_{i \in x} w_i s_i^+,$$

$$s^0 = \sum_{i \in x} w_i s_i^0,$$

$$s^- = \sum_{i \in x} w_i s_i^-$$

where x contains all the terms in a sentence and could be either of binary, term occurrence or tf-idf weights in the BOW representation of the sentence [1] . Now each sentence is represented as a 3-dim vector S as follows:

$$s = \left[s^+, s^0, s^-\right]^T$$

Support Vector Machines are supervised learning models used for classification.

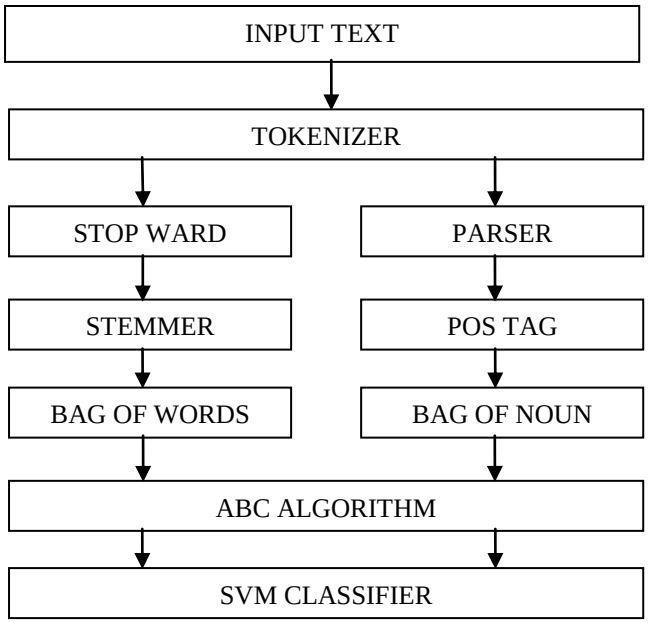

Fig 1: -Steps and techniques used in sentiment classification

E. Step used

Input Text: Input Text processing techniques are divided into two subcategories.
Tokenizer: Textual data comprises block of characters called tokens. The documents are separated as tokens and used for further processing.
Parser : POS tags are tags based on the grammatical role of the terms in the sentence (nouns, verb, adjective, adverb, etc).

Stop words : List of stop words, Some of the more frequently used stop words for English include "a", "of","the", "I", "it", "you", "and" and special characters like.,[]{}()/ these are generally regarded as 'functional words' which do not carry meaning. If the stop word removal is applied, all the stop words in the particular text file will not be loaded.
Stemmer: The stemming is the process for finding the root words or is the procedure of describing relevant tokens into a single type [13]. For example "He teach us in an interesting manner" This sentence after stemming is converted into "teach interest manner" thus, by using stem (root) word, the comparison of sentence word with number of positive/negative words becomes easy.
Bag of Words : The BOW list is constructed by considering all the words from sentiments sentence.
Bag of Noun : The BON list is constructed by considering only the nouns/noun phrases.
ABC Algorithm : The ABC algorithm was first applied to numerical optimization [20].
ABC algorithm is a new swarm intelligent algorithm, which proposed by Karabog in Erciyes University of Turkey in 2005 .Since ABC algorithm is simple in concept, easy to implement, and has fewer control parameters, it has been widely used in many optimization applications such as protein tertiary structures , digital IIR filters artificial neural networks [19].
The minimal model of foraging selection that leads to the emergence of collective intelligence of honey bee swarms consists of three essential components: food sources, employed foragers and unemployed foragers. There are two basic behaviours: recruitment to a food source and the abandonment of a food source [19].

1) Food sources: It represents the position of solution for optimization problem, the profitability of food source are expressed as fitness of the solution.

2) Unemployed foragers: they are two types, scouts and onlookers. Their main task is that exploring and exploiting food source. At the beginning, there are two choices for the unemployed foragers: (i). It becomes a scout – randomly search new food sources around the nest; (ii). It becomes an onlooker – determine the nectar amount of food source after watching the waggle dances of employed bee, and select food source according to profitability [19].

3) Employed foragers: the honeybees found food source, which also known as the employed bees, are equal to the number of food sources. The employed bees store the food source information and share with others according to a certain probability. The employed bee will become a scout when food source has been exhausted [19]. Basically, there are two important function supports the algorithm:

$$P_i = \frac{f_i t_i}{\left(\sum_{n=1}^{sn} f_i t_n\right)} \qquad \text{.....Eq. (4.1)}$$

$$V_{ij} = X_{ij} + \phi(X_{ij} - X_{kj}) \qquad \text{.....Eq. (4.2)}$$

Where Pi is the probability value associated with ith food source that calculated by the Eq.4.1. An onlooker bee selects a food source relying on Pi. In this equation, fiti represents ith food source's nectar amounts, which is measured by employed bees and SN is the number of food source which is equal to the number of employed bees [19].

In ABC algorithm, the artificial bees need to do the local search for finding the new possible solution. Eq.4.2 [19] shows the local search strategy of the original ABC optimization algorithm. Vij is the new candidate food position produced by this equation where k € [1,2,…,SN] and j € [1,2,…,D] are randomly chosen parameters, but k has to be different from j. Thus Xij and Xkj represent the different old food source positions. The difference between these two positions is the distance from one food source to the other one. SN is the number of employed bees as the previous described and D is the number of optimization parameters. Φij is a random number between [-1, 1] and controls the distance of a neighbour food source position around Xij [19].

The most crucial factor in ABC is the processing of greedy selection which means if the new food has equal or better nectar than the previous source; it will replace the previous one in the memory. Otherwise, the previous one is retained. In other words, a greedy selection mechanism is employed as the selection operation between the previous and

the current food sources [19].

Beside the SN control parameter, there are two other parameters used in the basic ABC algorithm: the limitation value of food source position and the maximum number of searching cycle. The users should carefully choose these two values, since the larger number will slow down the optimization process significantly, and the small value cannot find a good solution for the requirement [19].

SVM Classifier:SVM are often considered as the classifier that makes the greatest accuracy outcomes in the text classification issues [13]. SVM classifier first, index the term of opinion ascending. Then, all the term are weighted according to its features. If the Score of weighting is greater than zero, then term is classified as positive reviews, then term is classified as negative review[13].

V. EXPECTED RESULT AND DISCUSSION

The proposed software will be efficiently calculate the weight of sentiment. The software will give the better accuracy, result with weighted sentiment, improved classifier error rate on sentiment opinion.

Hence very limited work has been done the field of increasing the accuracy of the classification of sentiment opinion's or reviews. Optimal feature selection is used for reducing feature subset size and computational complexity thereby increasing the classification accuracy. The ABC algorithm being a powerful optimization technique and is widely used for solving combinatorial optimization problems. Hence, this method is incorporated for optimizing the feature subset selection in this investigation. . Experiment results evaluated feature selection techniques based on proposed ABC. Experimental results show that the classification accuracy of the classifiers improves feature selection by using ABC algorithm.

So we want to develop software that will improve the accuracy of classification by using ABC algorithm that reduce feature subset size and computation complexity and increase the classification accuracy.

VI. CONCLUSION

In this paper we studied the aspect level of sentiment analysis considering the three classes for sentiment polarity of sentence. The main contribution of this paper is the proposal of a feature set selection, weighted, that lead to more accurate sentiment analysis. This scheme is based upon the three scores (positiveness, neutralness and negativeness) that are learned from the data for each term. Optimal feature selection is used for reducing feature subset size and computational complexity thereby increasing the classification accuracy. The ABC algorithm being a powerful optimization technique and is widely used for solving combinatorial optimization problems. Hence, this method is incorporated for optimizing the feature subset selection in this investigation. Using sentiment analysis, features from text are extracted, and classified those providing opinions/sentiments about text/data/documents through Support Vector Machine classifier. The ABC algorithm used for feature selection to improve the accuracy of the classifier result of sentiments of opinion/sentiments.

REFERENCES

[1] Farhadloo.M, Rolland.E, "Multi-Class Sentiment Analysis with Clustering and Score Representation" 13th International Conference on Data Mining Wokshop,IEEE,2013.

[2] S.Njolstd.P, S.Hoysaeter.L, Wei.W and Atle Gull.J. "Evaluating Feature Sets and Classifiers for Semtiment Analysis of Financial News," IEEE/WIC/ACM International Joint Conferences on Web Intelligence(W1) and Intelligent Agent Technologies(IAT),IEEE,2014.

[3] Valakunde.N, Patwardhan.M,"Multi-Aspect and Mutli-Class Based Document Sentiment Analsis of Educational Data Caterin Accreditaion Process" International conference on Cloud & Ubiquitous Computing & Emerging Technologies,IEEE,2013.

[4] Liu.L, Nie.X, Wang.H,"Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis" 5th International Congress on Image and Signal Processing(CISP 2012),IEEE,2012.

[5] Sumathi.T, Karthik.S, Marikkannan.M, "Arificial Bee Colony Optimation For Feature In Opinion Mining ",Journal of Theoretical and Applied Information Technoloy, Elsevier, 2012.

[6] Basari.A, Hussin.B, Ananta.I, Junta Zeninarja,"Opinion Mining of Movie Review using Hybrid Method of support Vector Machine and Particle Swarm Optimization" Malaysian Technical Universites Conference on Engineering & Techonlogy 2012, MUCET 2012 Part 4-Information And Communication Technology, Elsevier, 2013.

[7] ManjuS.R, Kalaiman.E, R.Bhavani, "Product Aspect Ranking Using Sentimentic Oriented Sentiment Classifier" IJERA, 2014.

[8] Singh.N, Ghalib.M.R,"An Effective E-Commerce Managemnet using Mining Techniques" International Journal of scientific and Research Publictions, Volume 3, Issue 8, IJSPR,2013.

[9] R.Shikalgar.N, Badgujar.D,"Online Review Mining For Forcasting Sales" IJRET, 2013.

[10] Vigneshkumar K, Gnanavel S,"Mining Online Reviews for Predicting Sales Performance in Movie Domain" TIJCSA, 2013.

[11] S.Modha.J Prof & Head S.Pandi.G , J.Modha.S, "Automatic Sentiment Analysis for Unstructured Data", IJARCSSE,2013 .

[12] Patil.G, Galande,V, Kekam.V, Dange.K, "Sentiment Analysis Using Support Vector Machine" IJIRCCE,2014.

[13] Basiri.M.E, Naghsh-NilchiA.R, AND Ghasem-Aghaee.N,"Sentiment Prediction Based On Dempster-Shafer Theory Of Evidence" HINDAWI,2014.

[14] Tripathi.G, S.N,"Opinion Mining: A Review" IJICT,2014.

[15] Joshi.N,Itkat.S,"A Survey on Feature Level Sentiment " IJCSIT, 2014.

[16] Saif.H, He.Y and Alani.H,"Semantic Sentiment Analysis of Twitter", ISWC,2012.

[17] Patni.S, Wadhe.A, "Reviews Paper on Sentiment Analysis is - Big Challenge" IJARCSMS,2014.

[18] Varghese.R, "A Survey On Sentiment Analysis and opnion mining" association rules", IJRET, 2013.

[19] Shukran.M, Yeh.W, Wahid.N, Zaidi.A, "Artificial Bee Colony based Data Mining Algorithms for Classification Task" Vol.5,No 4,August 2011.

[20] Akay.B, Karaboga.D, "A modified Artifical Bee Colony algorithm for real-Parameter optimization " ELSEVIER, 2010.

[21] Murugan.R, Mohan.M, "MODIFIED ARTIFICIAL BEE COLONY ALGORITHM FOR SOLVING ECONOMIC DIPATCH PROBLEM" ARPN, 2012.

[22] Khaze.S, Maleki.I, Hojjatklhah.S, Bagherinia.A,"Evaluation The Efficiency of Artificiaal Bee Colony and The Fireefly Algorithm in Solving The Continuous Optimization Prpoblem " IJCSA, Vol.3,No.4,2013.

[23] Joshi.N, Itkat.S,"Feature Selection with Chaotic Hybrid Artifical Bee Colony Algorithm based on Fuzzy (CHABCF)" ISPACS, 2013.

About Authors:

Ms. Ruby Dhurve received the B.E. degree from Chhattigarh Swami Vivekanand Technical University, Bhilai (C.G.) India in Computer Science & Engineering in the year 2013. She is currently pursuing M.Tech. Degree in Computer Science Engineering with specialization in Computer Science & Engineering from CSVTU Bhilai (C.G.), India. Her research area includes Data Mining and Text Mining etc.

Ms. Megha Seth is currently Assistant professor in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. She completed her B.E and M.Tech. in Computer Science and Engineering Branch , experience nine year read RCET Bhilai. Her research area includes Data mining, Image processing, Computer Network, AI & NN etc. She has published many Research Papers in various reputed National & International Journals,

