Journal of Theoretical and Applied Information Technology
10th August 2014. Vol. 66 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ARTIFICIAL BEE COLONY OPTIMIZATION FOR FEATURE SELECTION IN OPINION MINING

[1]T. SUMATHI, [2]S.KARTHIK, [3]M.MARIKKANNAN
[1 & 3] Department of Computer Science and Engineering, Institute of Road and Transport Technology, Erode, Tamilnadu, India.
[2] Department of Computer Science and Engineering, SNS College of Technology, Coimbatore, Tamilnadu, India
E-mail: [1]sumathi.opm@gmail.com

ABSTRACT

Opinion mining, a sub-discipline of information retrieval and computational linguistics concerns not with what a document is about, but with its expressed opinion. Feature selection is an important step in opinion mining, as customers express product opinions separately according to individual features. Earlier research on feature-based opinion mining had many drawbacks like selecting a feature considering only grammatical information or treating features with same meanings as different. However this led to a large corpus which subsequently affected the classification accuracy. Statistical techniques like Correlation Based Feature (CFS) have been extensively used for feature selection to reduce the corpus size. The selected features are sub optimal due to the Non Polynomial (NP) hard nature of the technique used. In this work, we propose Artificial Bee Colony (ABC) algorithm for optimization of feature subset. Naïve Bayes, Fuzzy Unordered Rule Induction Algorithm (FURIA) and Ripple Down Rule Learner (RIDOR) classifiers are used for classification. The proposed method is compared with features extracted based on Inverse Document Frequency (IDF). Hence, this method is useful for reducing feature subset size and computational complexity thereby increasing the classification accuracy.

Keywords: *Opinion mining, Feature selection, Artificial Bee Colony (ABC), IMDb dataset, Inverse Document Frequency (IDF), Naive Bayes, Fuzzy Unordered Rule Induction Algorithm (FURIA), Ripple Down Rule Learner (RIDOR)*

1. INTRODUCTION

Opinion mining has recently attracted attention within fields such as marketing, personal affective profiling, and financial market prediction which is a computational linguistics sub discipline concerned with the opinion a topic expresses [1]. It is defined as computational study of opinions, sentiments and emotions in texts [2]. Opinions exist for any entity or object (person, product, service, etc.) on the Web and their features/components like cell phone battery, keyboard and touch screen display. Opinion mining tasks can be converted to classification tasks, to enable machine learning use for opinion mining.

Opinion mining is a research domain that deals with automatic detection and extraction methods of opinions and sentiments in a text [3]. User's opinions and sentiments are easy to extract when applied to an entity (product, film). Detection process is easy at sentence level and for document level by aggregating individual results, and applying a specific algorithm. Opinion mining is also called Sentiment analysis which attempts to identify and analyze opinions or emotions [4]. Sentiment analysis, deals with direction-based text analysis e.g. text having opinions and emotions.

Feature selection select relevant features based on a specific measurement, its purpose being to simplify training and reduce training time [5]. A feature is considered relevant to specific class if its existence in a particular class is high compared to its existence in another class. Though text opinion mining involves important tasks, assigning sentiment polarities accurately (like positive, negative, neutral) and intensities (like high or low) is still a challenge [6].

Feature selection consists of four steps namely subset generation, subset evaluation, stopping criterion and result validation. Subset generation is defined as search procedure for evaluation of candidate feature subsets based on

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

specific search strategy. Candidate subsets are evaluated and compared to a previous best based on specific evaluation criteria. The better and new subset replaces the previous one of similar quality. This subset generation and evaluation is repeated until a specific stopping criterion is satisfied. The selected best subset is then validated by earlier knowledge or different tests through synthetic and/or real world data sets.

Features denote properties of textual data in text classification, and are measured to classify text, like bag-of words, n-grams (unigram, bi-grams, tri-grams), header information, word position and ordered word list [7]. Present feature selection procedures in machine learning like Document Frequency (DF), Chi Square and Information Gain assign values to features based on specific statistical equation. Features are sorted and then it is for the user to select appropriate features based on sorted value [8]. Different users select varied features. Usually, a newbie unaware of this scenario does nothing causing classifier transaction to consume more processing time using more resources.

Feature selection fall into two categories, filters and wrappers. The former ranks features by valuation criterion retaining only features with values above a threshold [9]. Wrapper methods search features set for optimal subsets in a specific classifier. Performance measures are attached to subsets based on its performance using a specific learning dataset classifier.

Original features subsets are selected by feature selection, and feature subsets optimality is measured through evaluation criterion. When domain dimensionality expands, N features number increases. This is intractable to locate an optimal feature subset. Feature selection related problems are NP-hard. Machine learning feature selection is a global optimization problem, that reduces features number, removes irrelevant, noisy and redundant data, resulting in recognition accuracy [10]. It is an important step affecting pattern recognition system performance. Usually, feature selection problems are solved through use of single objective optimization techniques like genetic algorithm [11]. Such techniques only optimize a single quality measure; for e.g., recall, precision or F-measure at a time. Sometimes, a single measure is unable to capture a good classifier's quality reliably. A good classifier must have recall, precision and F-measure values optimized simultaneously instead of any parameter high value alone.

Swarm intelligence is widely used to address selection of optimal feature set. An Artificial Bee Colony (ABC) algorithm, a recently introduced optimization algorithms, to simulate intelligent foraging behaviour of honey bee swarm was proposed by Karboga and Ozturk [12]. Clustering analysis is used in many disciplines/applications to identify homogeneous groups of objects based on their attributes values. ABC was used for data clustering on benchmark problems and its performance was compared to particle swarm optimization algorithm and nine other classification techniques from literature. Thirteen typical test data sets from UCI Machine Learning Repository demonstrated techniques results. Simulation indicated that ABC algorithm was efficient for multivariate data clustering. An ABC algorithm which is an optimization algorithm based on honeybee swarms intelligent behaviour was proposed by Karboga and Basturk [13]. This compares performances of ABC algorithm with differential evolution, particle swarm optimization and evolutionary algorithm for multi-dimensional numeric problems. Simulation results revealed that ABC algorithm's performance was comparable to the above mentioned algorithms and can be used to solve high dimensionality engineering problems.

This study uses ABC optimization algorithm for feature selection. The proposed method extracts a feature set based on Inverse Document Frequency (IDF). The selected features are classified using Naive Bayes, RIDOR and FURIA. IMDb, a movie database is used. Rest of the paper is organized as follows: Section 2 reviews work in literature. Section 3 describes methods used while section 4 discusses about experiments and results. Section 5 concludes the study.

2. RELATED WORKS

A comparative study on methods and resources to be used for mining opinions was presented by Balahur et al [14]. This tasks' difficulty, motivated by presence of various possible targets affect phenomena that quotes contain. The evaluation of the approaches is through methods which uses annotated quotations from news provided by EMM news gathering engine. It concludes that generic opinion mining systems require both use of large lexicons and specialized training/testing data.

Techniques and approaches that directly enable opinion-oriented information-seeking systems were presented by Bo et al [15]. This focused on methods seeking to address new challenges raised by sentiment-aware applications, compared to those

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

already presented in traditional fact-based analysis. The material summarized evaluative text and broader issues regarding privacy, manipulation, and economic impact that opinion-oriented information-access services development resulted in. To ensure future work, a discussion of available resources, benchmark datasets, and evaluation campaigns was provided. With availability and popularity of opinion-rich resources like online review sites and personal blogs, new opportunities and challenges come up, due to people now using information technologies to understand others opinions. The sudden activity in opinion mining and sentiment analysis deals with computational treatment of opinion, sentiment, and text subjectivity.

Improving aspect-level opinion mining for online customer reviews was focused on by Xu et al [16] who first proposed a new generative topic model, the Joint Aspect/Sentiment (JAS) model, to extract aspects and aspect-dependent sentiment lexicons from online customer reviews. An aspect-dependent sentiment lexicon refers to aspect-specific opinion words with aspect-aware sentiment polarities regarding specific aspects. Then extracted aspect-dependent sentiment lexicons are applied to series of aspect-level opinion mining tasks, including aspect identification, aspect-based extractive opinion summarization, and aspect-level sentiment classification. Experiments demonstrated the JAS model's effectiveness in learning aspect-dependent sentiment lexicons and practical values of extracted lexicons applied to such practical tasks.

Advent of Web 2.0 and social media content which stirred excitement and created abundant opportunities to understand opinions of the public/consumers to political movements, social events, marketing campaigns, company strategies and product preferences was proposed by Zhang et al [17]. Many exciting social, geopolitical, and business-related questions are answered by analyzing thousands, and millions of comments/responses in various blogs (blogosphere), fora (Yahoo Forums), social media and social network sites (YouTube, Facebook, and Flikr), virtual worlds (Second Life), and tweets (Twitter). Opinion mining, a sub discipline in data mining and computational linguistics refers to computational techniques to extract, classify, understand, and assess opinions in various online news sources, social media comments, and user-generated content. Sentiment analysis is used in opinion mining to identify sentiment, affect, subjectivity, and other online text emotional states.

A model first identifying product feature before collecting their positive and negative opinions to produce a summary of good/bad points was proposed by Zhai et al, [18]. Review aggregators and e-commerce sites are just examples of businesses reliant on opinion mining to produce feature-based products' quality summaries.

A systematic evaluation of feature selectors and feature weights with Naive Bayes and Support Vector Machine classifiers was proposed by O'Keefe and Koprinska [19] including introduction of two new feature selection methods and three feature weighting methods. Sentiment analysis identifies whether opinion in a document is positive/negative on a topic. Many sentiment analysis applications are presently infeasible due to huge number of features in corpora. Results revealed it was possible to maintain an 87.15%, state-of-the art classification accuracy when using less than 36% features.

Use of sentiment analysis methods to classify web for opinions in multiple languages was proposed by Abbasi et al [20]. Stylistic and syntactic features utility was evaluated for sentiment classification of English and Arabic content. Specific feature extraction components are integrated to account for the Arabic language's linguistic characteristics. The Entropy Weighted Genetic Algorithm (EWGA) a hybridized genetic algorithm incorporating information gain heuristic for feature selection was developed. EWGA aimed to improve performance and ensure key features better assessment. The proposed features were evaluated on a benchmark movie review data set and U.S. and Middle Eastern web fora postings. Experimental results using EWGA with Support Vector machine (SVM) reveal high performance levels with more than 95% accuracy on benchmark data set and over 93% for U.S. and Middle Eastern fora. Stylistic features enhanced performance across test beds while EWGA outperformed other feature selection procedures indicating these features and techniques utility for document level sentiment classification.

Unstructured text being the primary means for publishing biomedical research results was proposed by Swaminathan et al [21]. To extract and integrate such data's knowledge, text mining was applied routinely. An important task was extracting relationships between bio-entities like foods and diseases. Present studies stop short of analyzing extracted relationships like polarity and certainty level at which authors reported a relationship. The latter was termed relationship strength and marked

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

at three levels; weak, medium and strong. A preliminary study was previously reported on this issue and the studies are detailed on constructing a new feature space to effectively predict a relationship's polarity and strength.

An enhanced feature extraction and refinement method called FEROM to extract correct features from review data through exploitation of grammatical properties and semantic characteristics of feature words was proposed by Jeong et al [22]. It refines features by recognizing and merging those which are similar. Opinion mining analyzes customer's opinions using product reviews providing meaningful information including opinions polarity. In opinion mining, feature extraction is important as customers express product opinions separately according to individual features. Experiments performed on actual online review data proved that FEROM was effectively extracted and refined features to analyze customer review data and eventually contributed accurate and functional opinion mining.

A range of feature selectors was systematically evaluated with regard to efficiency in improving classifiers performance for sentiment analysis [23]. Sentiment analysis used movie reviews. Work to the main subtask of opinion summarization was proposed by Somprasertsri and Lalitrojwong [24]. Product feature and opinion extraction is important for opinion summarization, as its effectiveness greatly affects opinion orientation identification performance. It is important to identify semantic relationships between product features and opinions correctly.

Features based on syntactic dependency relations to be used to improve opinion mining performance were proposed by Joshi and Penstein-Rosé [25]. Using a dependency relation triples transformation; convert features into "composite back-off features" generalizing better than regular lexicalized dependency relation features. Experiments compared this approach with many other approaches which generalize dependency features or n-grams demonstrate of composite back-off features utility.

A novel co-occurrence association-based method aiming to extract implicit features in customer reviews providing comprehensive and fine-grained mining results was proposed by Zhang and Zhu [26]. Samsudin et al [27] used an artificial immune system based feature selection technique to select appropriated opinion mining features. Experiments using 2000 online movie reviews proved that the technique reduced 90% features and improved opinion mining accuracy up to 15% and up to 6% with k Nearest Neighbour classifier and Naive Bayes classifier respectively.

An Accelerated Artificial Bee Colony (A-ABC) method is suggested by Ozkis and Babalik [28] where two modifications were used on ABC algorithm to ensure local search ability and convergence speed. Modifications were called Modification Rate (MR) and Step Size (SS). Performances of A-ABC and standard ABC were compared with 7 different benchmark functions to investigate effects of using MR value and SS modification. Results showed that A-ABC generally performed better and had faster convergence than ABC algorithm's standard version.

Analysis of linguistic resources required for opinion mining and a few machine learning techniques based on their usage and importance for analysis and evaluation of Sentiment classifications and applications was surveyed by Padmaja and Fatima [29]

A Modified Artificial Bee Colony (MABC) algorithm applied to GMS optimization problem efficiently was proposed by Anandhakumar et al [30]. This algorithm was proposed to handle system constraints effectively and to ensure better maintenance schedules. The proposed algorithm's efficacy was illustrated with 13 generating units and 21 generating units with two differing load demands. Simulation results were compared to discrete particle swarm optimization, modified discrete particle swarm optimization and multiple swarms - modified discrete particle swarm optimization all population based heuristic search algorithms. It was seen from the numerical results, that MABC based approach ensured better solutions for GMS.

A new feature selection method using ABC algorithm to optimize feature selection was proposed by Palanisamy and Kanmani [31]. Ten UCI datasets were used to evaluate the proposed algorithm. Experimental results showed that ABC-feature selection resulted in optimal feature subset configuration and achieved a classification accuracy of 98.55%. The proposed method increased classification accuracies by 12% compared to classifier/standard ensembles such as J48, bagging and boosting of C4.5.

An ABC algorithm, a recently proposed algorithm, was tested on fuzzy clustering, and used to classify different data sets by Karboga and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Ozturk [32]; a collection of classification benchmark problems including Cancer, Diabetes and Heart from UCI database were used. Results indicate that ABC optimization algorithm performance was successful in achieving lower classification error percentage of 16.32%. The proposed ABC fuzzy clustering achieves 8.09% less classification error when compared to Fuzzy C-Means.

3. METHODOLOGY

A. Internet Movie Database (IMDb)

Original data source is the Internet Movie Database which consists of relevant and comprehensive information about past, present and future movies. It started as a set of shell scripts and data files [33]. The IMDb network is bipartite (two mode) having 1324748 = 428440+896308 vertices and 3792390 arcs. 9927 arcs in network are multiple (parallel) arcs. IMDb uses two methods of adding information to database: Web forms and e-mail forms.

Information from a submission procedure reveals it is easier to use web forms rather than humungous e-mail format, if addition to information is an update [34]. If altogether new information is submitted, a user requests and obtains IMDb format templates via e-mail. The information to be submitted has to attest according to templates and validated. Information submitted for a title at IMDb includes five sections: The Title, Production Status, Cast, Crew, and Miscellaneous. This study uses ABC optimization algorithm for feature selection, and experiments were conducted using classification models based on Naive Bayes, RIDOR and FURIA.

Stop words are function words like prepositions, articles, conjunctions and pronouns, providing language structure rather than content [35]. Such words do not impact category discrimination. Additionally, common words like ‘a’ and ‘of’, may be removed as they occur frequently so as not to be discriminating for a specific class. Common words are identified by a threshold on number of documents the word occurs in, e.g. if it occurs in over half of documents, or by supplying a stop word list. Stop words are language and domain-specific. Depending on classification task, they can risk removing words which are essential predictors, e.g. the word ‘can’ is discriminating between ‘aluminium’ and ‘glass’ recycling.

Word stemming is a crude pseudo-linguistic process that removes suffixes to reduce words to word stem. For example, words ‘classifier’, ‘classified’ and ‘classifying’ can be reduced to word stem ‘classify’. The common practice of stemming or lemmatizing—merging various word forms like plurals and verb conjugations into a distinct term— reduces features number to be considered. Properly, it is a feature engineering option.

B. Inverse Document Frequency (IDF)

The term frequency and Inverse Document Frequency (IDF) [36] help to identify a feasible collection of product features. IDF is a numerical statistic reflecting how important a word is to a document, in a collection/corpus. It is often a weighting factor in information retrieval and text mining. IDF value increases proportionate to a word appears many times in a document, but is offset by word frequency in the corpus, which helps control the fact that some words are more common than others. IDF weighting scheme variations are used by search engines in scoring and ranking a document's relevance from a given user query. IDF can be used for stop-words filtering in different subject fields like text summarization and classification. Text Classification is a semi-supervised machine learning task automatically assigning a given document to a pre-defined categories set based on textual content and extracted features.

$$IDF(a) = \log \frac{1+|x|}{x_a} \quad (1)$$

where, x_a is the set of documents containing term a.

It is usual to replace simple word frequency with a weighted frequency before computing cosine and other statistics. Weighted frequency statistic is TFIDF (Term Frequency Inverse Document Frequency) statistic which computes a weight for every term reflecting its importance. The term’s relevance to a specific document depends on how many words it has. This is why TFIDF denominator is adjusted for words number in a document. It is also adjusted for number of records (or documents) having the word (terms appearing on many records are down-weighted) [37].

The TFIDF formula is:

$$TFIDF = \frac{Frequency(i) * N}{df(i)} \quad (2)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where,

df is the frequency of word (i) in all documents

N is the number of words in the record /document

i is the word list in record/document.

In TFIDF representation, term frequency for every word is normalized by IDF [38] which reduces weight of terms occurring frequently in a collection. This also reduces the importance of common terms in a collection, ensuring that document matching is influenced by more discriminative words with relatively low frequencies in a collection.

C. The Naive Bayes Classifier

The Naive Bayes Classifier is a popular algorithm due to its simplicity, computational efficiency and performance for real-world problems. It is a supervised learning technique and also a statistical technique for classification [39]. This method assumes an underlying probabilistic model and allows for capture of uncertainty about the model in a principled way through determining outcome probabilities.

It is a text classification that assigns classc*=argmaxcP(c|d), to a given document d. A Naive Bayes classifier is a probabilistic classifier based on Baye's theorem and is suited when inputs dimensionality is high. Its underlying probability model is described as an "independent feature model". The Naive Bayes (NB) classifier uses Bayes' rule Equation [40],

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \qquad (3)$$

where, $P(d)$ plays no role in selecting c*. To estimate term $P(d/c)$, Naive Bayes decomposes it by assuming f_i's are conditionally independent given d's class as in Equation No.(4).

$$P_{NB}(c \mid d) = \frac{P(c)(\prod_{i=1}^{m} p(f_i \mid c)^{n_i(d)})}{P(d)} \qquad (4)$$

where, m is the number of features and f_i is the feature vector.

Uses of Naive Bayes classification [41]:

- The Bayesian classification is a probabilistic learning method which is used for classifying text documents.
- Spam filtering is best known use of Naive Bayesian text classification. It uses a Naïve Bayes classifier to identify spam e-mail and to distinguish illegitimate spam email from the legitimate.
- Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering Recommender Systems use machine learning and data mining techniques to filter unseen information and predict if a user can be a given resource.

D. Fuzzy Unordered Rule Induction Algorithm (FURIA)

FURIA is an inverse solution-based algorithm that learns and uses subject-specific features for mental state classification. A feature extracted through FURIA corresponds to activity in a ROI and its associated frequency band. Compared to current methods, FURIA automatically identifies relevant ROI, and also frequency bands where ROI current densities are discriminant [42].

Finally, FURIA introduces concepts of fuzzy ROI and fuzzy frequency bands to get increased classification performances. FURIA aims to be modular in that various inverse solutions are possible within it. This section describes inverse solutions that are possible within FURIA and a specific one which is used in implementation. It also describes the FURIA feature extraction algorithm.

E. Ripple Down Rule Learner (RIDOR)

RIDOR is an abbreviation for Ripple Down Rule learner which generates a default rule first and then exceptions for it with least (weighted) error rate. It then generates "best" exceptions for every exception and iterates until it is pure. It thus performs a tree like exceptions expansion and the leaf has only default rules without exceptions. Exceptions are rules set which predict a class other than class in a default rule. IREP detects exceptions [43].

F. Artificial Bee Colony (ABC)

This study proposes a wrapper using ABC algorithm. The ABC algorithm is population-based inspired by bees foraging for food. The algorithm is divided into 2 groups comprising worker and non-worker bees. Non-worker bees include onlooker bees and scout bees. In the algorithm, all worker bees are sent to half the food source. After calculation of food amount at a source, a local

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

search is implemented around food by onlooker bees. If another source with additional nutrition is explored, it is updated as a new food source [44]. If a better food source is not found by onlooker or scout bees following a certain cycle, onlooker bees are oriented to a random food source. With each cycle, bees in the population are sorted according to food sources they locate and best bees are transferred to next cycle. When stopping criterion is ensured, the algorithm ends. The initial points for the algorithm and food sources that worker bees will fly to, are determined by Equation,

$$u_{ij} = u_{\min j} + rand \times (u_{\max j} - u_{\min j}) \quad (5)$$

where,

$u_{\min j}$ and $u_{\max j}$ show the minimum and maximum of the variable u_j .

Each search cycle has three steps after initialization: moving employed bees to food sources and calculating nectar amounts; placing onlookers onto food sources and calculating nectar amounts; determining scout bees and directing them to possible food sources [45]. A food source represents a possible solution to the problem being optimized. The amount of nectar in a food source corresponds to solution quality represented by that food source.

The flowchart for the proposed system is shown in figure 1. The unstructured data is prepared for processing by computing the word score using correlation technique. The ABC method is initiated to compute the fitness value. This process has been iterated to obtain the efficient classification accuracy.

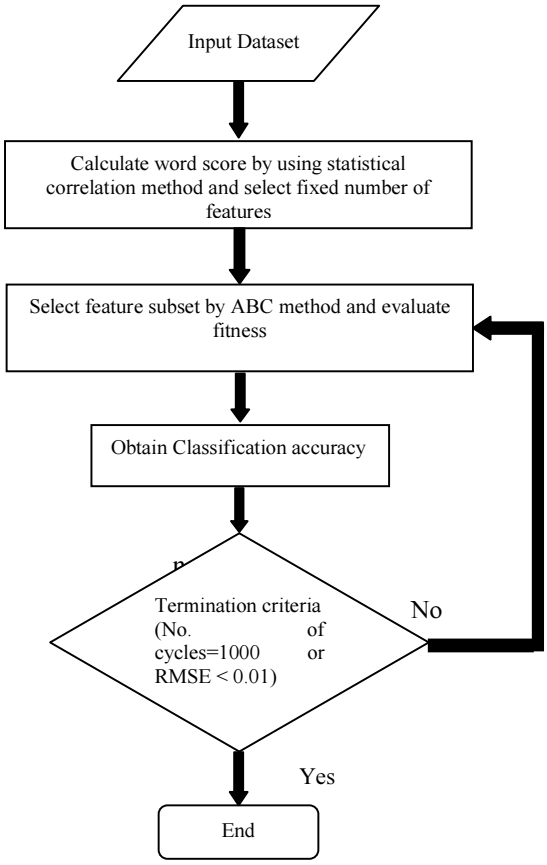

Figure 1: *The flow chart of the proposed system*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

G. Algorithm for ABC optimization

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 (binary bit string) to each employed bee
      ii.Produce new feature subsets v_i
      iii.Pass the produced feature subset to the classifier
      iv.Evaluate the fitness (fit_i) of the feature subset
      v.Calculate the probability p_i of feature subset solution
   b. Construct solutions by the onlookers
      i.Select a feature based on the probability p_i
      ii.Compute v_i using x_i and x_j
      iii.Apply greedy selection between v_i and x_i
   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.

4. EXPERIMENTS AND RESULTS

In our previous work [46], performance analysis for classification methods for opinion mining features are selected using proposed optimized method. The features are classified using Naïve Bayes, FURIA and RIDOR. The classification accuracy, Root Mean Square Error (RMSE), Precision and Recall are calculated as follows:

$$\text{Classification accuracy } (\%) = \frac{(TN + TP)}{(TN + FN + FP + TP)} \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|e_i|^2} \quad (7)$$

where, n is the number of instance and e_i is the error.

$$\text{Precision} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FP} \quad (9)$$

Journal of Theoretical and Applied Information Technology
10th August 2014. Vol. 66 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where,

TN (True Negative) - Number of correct predictions that an instance is invalid

FP (False Positive) - Number of incorrect predictions that an instance is valid

FN (False Negative)- Number of incorrect predictions that an instance is invalid

TP (True Positive) - Number of correct predictions that an instance is valid

The table 1 shows the results obtained for classification accuracy.

Table 1 Classification accuracy

Techniques used	Features - IDF	Features – ABC
Naïve Bayes	85.25	88.5
RIDOR	76	78.5
FURIA	92.25	93.75

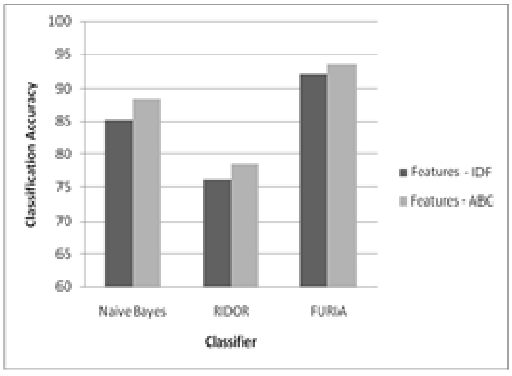

Figure 2: Classification accuracy

From the figure 2, it is observed that proposed ABC for feature selection improves classification accuracy by 1.63% to 3.81% when compared with IDF. The FURIA method achieves the classification accuracy of 93.75% and is higher when compared with RIDOR by 16.26% and by 5.6% for Naive Bayes.

Table 2 RMSE

Techniques used	Features - IDF	Features – ABC
Naïve Bayes	0.3764	0.3124
RIDOR	0.4899	0.4637
FURIA	0.2312	0.1906

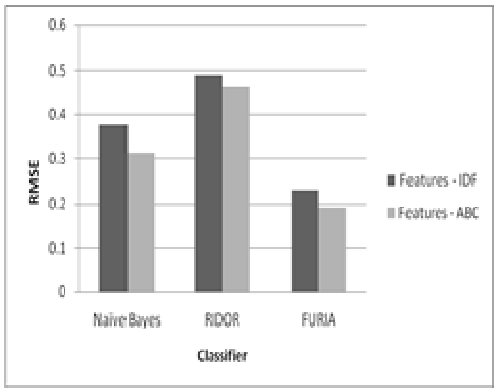

Figure 3: RMSE

From the figure 3, it is observed that RMSE is least for FURIA and the proposed feature selection reduces the RMSE by 5.35% to 17.56% for the classifiers.

Table 3 Precision

Techniques used	Features - IDF	Features – ABC
Naïve Bayes	0.853	0.887
RIDOR	0.764	0.785
FURIA	0.926	0.938

Journal of Theoretical and Applied Information Technology
10th August 2014. Vol. 66 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

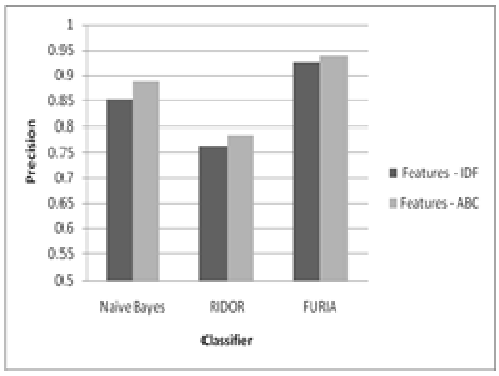

Figure 4: Precision

From the figure 4, it is observed that the precision for FURIA is better than RIDOR by 16.31% and Naïve Bayes by 5.43 %. The proposed ABC feature selection improves the precision by 1.3% to 3.99% when compared to IDF.

Table 4 Recall

Techniques used	Features - IDF	Features – ABC
Naïve Bayes	0.853	0.885
RIDOR	0.76	0.785
FURIA	0.923	0.938

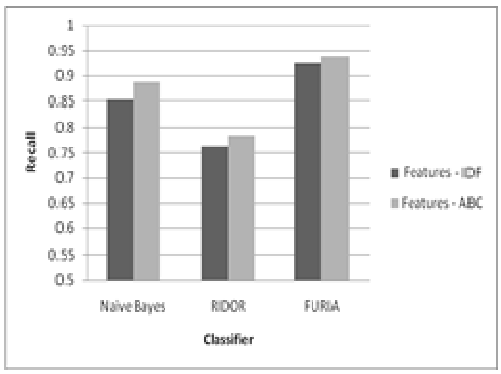

Figure 5: Recall

From the figure 5, it is observed that the proposed ABC feature selection method improves the recall by 1.63% to 3.75%. It shows that recall for FURIA is better than RIDOR and Naive Bayes by 16.31% and 5.65 % respectively.

5. CONCLUSION

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. For evaluation, IMDb movie dataset is used. The reviews are pre-processed by stemming and removal of stop words. Using sentiment analysis, features from text are extracted, and classified those providing opinions/sentiments about text/data/documents through Naive Bayes classifier, FURIA and RIDOR. 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 method.

REFRENCES:

[1] Cambria, E., Speer, R., Havasi, C., a Hussain, A. “Senticnet: A publicly available semantic resource for opinion mining”, Assoc. Advancement of Artificial Intelligence, pp.14-18, 2010.

[2] Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., and Gordon, J. “ Empirical study of machine learning based approach for opinion mining in tweets”, In Proceedings of the 11th Mexican international conference on Advances in Artificial Intelligence Springer Berlin Heidelberg. Vol.7629, pp. 1-14, 2013.

[3] Smeureanu, I., and Bucur, C. “Applying Supervised Opinion Mining Techniques on Online User Reviews”, Informatica Economica, Vol.16, No.2, pp.81-91, 2012.

[4] Abbasi, A., Chen, H., and Salem, A. “Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums”, ACM Transactions on Information Systems, vol.26, No.3, pp.12, 2008.

[5] Samsudin, N., Puteh, M., Hamdan, A. R., and Nazri, M. Z. A. “Immune Based Feature Selection for Opinion Mining”, In Proceedings of the World Congress on Engineering, Vol. 3,pp.1520-1525,2013.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[6] Chen, H., and Zimbra, D. "AI and opinion mining", Intelligent Systems, IEEE, Vol.25, No.3, pp.74-80, 2010.

[7] Chaovalit, P., and Zhou, L. "Movie review mining: A comparison between supervised and unsupervised classification approaches", In proceedings of IEEE 38th Annual Hawaii International Conference In System Sciences, pp. 112-113, 2005.

[8] Samsudin, N., Puteh, M., Hamdan, A. R., and Nazri, M. Z. A. "Mining Opinion in Online Messages", International Journal of Advanced Computer Science and Applications. Vol. 4, No. 8, pp.19-24, 2013.

[9] Sindhu, C., and ChandraKala, s. "A survey on opinion mining and sentiment polarity classification", International Journal of Emerging Technology and Advanced Engineering, Vol.3, No.1, pp.531-539, 2013.

[10] Ramadan, R. M., and Abdel-Kader, R. F. "Face recognition using particle swarm optimization-based selected features", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol.2. No.2, pp.51-65, 2009.

[11] Ekbal, A., Saha, S., and Garbe, C. S. "Feature selection using multi objective optimization for named entity recognition" In proceedings of IEEE 20th International Conference on Pattern Recognition, pp. 1937-1940,2010.

[12] Karaboga, D., and Ozturk, C. "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Applied Soft Computing, Vol.11, No.1, pp.652-657, 2011.

[13] Karaboga, D., and Basturk, B. "On the performance of artificial bee colony (ABC) algorithm", Applied soft computing, Vol.8, No.1, pp.687-697, 2008.

[14] Balahur, A., Steinberger, R., Goot, E. V. D., Pouliquen, B., and Kabadjov, M. "Opinion mining on newspaper quotations", In Web Intelligence and Intelligent Agent Technologies, Vol. 3, pp. 523-526,2009.

[15] Bo, Pang., and Lillian, Lee. "Opinion Mining on Newspaper Quotations", Foundations and Trends in Information Retrieval, Vol.2, No.1-2, pp. 1–135, 2008.

[16] Xu, X., Cheng, X., Tan, S., Liu, Y., and Shen, H. "Aspect-level opinion mining of online customer reviews", Communications, China, Vol.10, No.3, pp.25-41, 2013.

[17] Zhang, Y., Yu, X., Dang, Y., and Chen, H. "An Integrated Framework for Avatar Data Collection from the Virtual World: A Case Study in Second Life", IEEE Transaction on Intelligent Systems, Vol.25, No.6, pp.17-23, 2010.

[18] Zhai, Z., Liu, B., Wang, J., Xu, H., and Jia, P. "Product Feature Grouping for Opinion Mining", IEEE Transaction on Intelligent Systems, Vol.27, No.4, pp.37-44, 2012.

[19] O'Keefe, T., and Koprinska, I. "Feature selection and weighting methods in sentiment analysis", In Proceedings of 14th Australasian Document Computing Symposium, pp-67-74, 2009.

[20] Abbasi, A., Chen, H., and Salem, A. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", ACM Transactions on Information Systems, Vol.26, No.3, 2008.

[21] Swaminathan, R., Sharma, A., and Yang, H. "Opinion mining for biomedical text data: Feature space design and feature selection", In Proceedings of the 9th International Workshop on Data Mining in Bioinformatics, 2010.

[22] Jeong, H., Shin, D., and Choi, J. "Ferom: Feature extraction and refinement for opinion mining", ETRI Journal, Vol.33, No.5,pp.720-730, 2011.

[23] Isabella, J., and Suresh, R. "Analysis and evaluation of Feature selectors in opinion mining", Indian Journal of Computer Science and Engineering, Vol. 3 No.6, pp.757-762, 2013.

[24] Somprasertsri, G., and Lalitrojwong, P. "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization", Journal of Universal Computer Science, Vol.16, No.6, pp.938-955,2010.

[25] Joshi, M., and Penstein-Rosé, C. "Generalizing dependency features for opinion mining", In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp.313-316, 2009.

[26] Zhang, Y., and Zhu, W. "Extracting implicit features in online customer reviews for opinion mining", In Proceedings of the 22nd international conference on World Wide Web companion, pp. 103-104, 2013.

[27] Samsudin, N., Puteh, M., Hamdan, A. R., and Nazri, M. Z. A. "Mining Opinion in Online Messages", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 8, pp.19-24, 2013.

[28] Ozkis, A., and Babalik, A. "Accelerated ABC (A-ABC) Algorithm for Continuous

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Optimization Problems", Lecture Notes on Software Engineering, Vol.1, No.3, pp.262-266,2013.

[29] Padmaja, S., and Fatima, S. S. "Opinion Mining and Sentiment Analysis–An Assessment of Peoples' Belief: A Survey", International Journal of Ad hoc, Sensor & Ubiquitous Computing, Vol.4, No.1, pp.21-33, 2013.

[30] Anandhakumar, R., Subramanian, S., and Ganesan, S. "Modified ABC Algorithm for Generator Maintenance Scheduling", International Journal of Computer Science and Electrical Engineering, Vol.3, No.6, pp.812-819, 2011.

[31] Palanisamy, S., and Kanmani, S. "Artificial Bee Colony Approach for Optimizing Feature Selection", International Journal of Computer Science, Vol.9, No.3, pp. 432-438, 2012.

[32] Karaboga, D., and Ozturk, C. "Fuzzy clustering with artificial bee colony algorithm", Scientific research and Essays, Vol.5, No.14, pp.1899-1902, 2010.

[33] Ahmed, A., Batagelj, V., Fu, X., Hong, S. H., Merrick, D., and Mrvar, A. "Visualisation and analysis of the Internet movie database", In IEEE 6th International Asia-Pacific Symposium on Visualization, pp. 17-24, 2007.

[34] Avancha, S., Kallurkar, S., and Kamdar, T. "Design of Ontology for The Internet Movie Database (IMDb)", Semester Project, CMSC 771, Spring, pp.1-9, 2001.

[35] A, Mountassir., Houda, B., and Berrada., I. "A novel model for text document representation: application on Opinion mining datasets", International Journal of Computer Science Engineering and Information Technology Research, Vol.3, No.3, pp. 293-304,2013.

[36] Ahmad, T., Doja, M. N., and Islamia, J. M. "Ranking System for Opinion Mining of Features from Review Documents", International Journal of Computer Science, Vol.9, No.4, pp.440-447, 2012.

[37] Francis, L., and Flynn, M. "Text mining handbook. In Casualty Actuarial Society E-Forum", spring, pp.1-61, 2010.

[38] Aggarwal, C. C. "Mining text data", Springer Science, Business Media, 2012.

[39] Anand M., Anjali D., Amulya R., and Clayton G. "Opinion Mining for Text Classification", International Journal of Scientific Engineering and Technology, Vol.2, No.6, pp. 589-594, 2013.

[40] Sindhu C., and Dr. S. ChandraKala, "Opinion mining and sentiment classification: a survey", Ictact journal on soft computing, Vol.3, No.1, pp.420-427, 2012.

[41] Lotte, F., Lécuyer, A., and Arnaldi, B. "FuRIA: an inverse solution based feature extraction algorithm using fuzzy set theory for brain–computer interfaces", Signal Processing, IEEE Transactions on, Vol.57, No.8, pp.3253-3263, 2009.

[42] Novakovic, J., Minic, M., and Veljovic, A. "Genetic Search for Feature Selection in Rule Induction Algorithms", 18th Telecommunications forum TELFOR, pp.1109-1112, 2010.

[43] Kursat, A., and Ulas, K. "Solution of transient stability-constrained optimal power flow using artificial bee colony algorithm", Turkish Journal of Electrical Engineering & Computer Sciences, Vol.21, pp. 360-372, 2013.

[44] Alatas, B. "Chaotic bee colony algorithms for global numerical optimization", Expert Systems with Applications, Elsevier, Vol. 37, No.8, pp. 5682–5687, 2010.

[45] Shanmugapriya, P., and Kanmani, S. "Artificial Bee Colony Approach for Optimizing Feature Selection", International Journal of Computer Science Issues, Vol.9, No.3, pp. 432-438, 2012.

[46] Sumathi, T., Karthik, S., and Marikannan, M. "Performance Analysis of Classification Methods for Opinion Mining", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2, No.4, pp.171-177, 2013.