The International Journal of Engineering and Science (IJES)
|| Volume || 6 || Issue || 9 || Pages || PP 61-72 || 2017 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Feature Selection Using Hybrid Approach for Opinion Spam Detection

G. Janani, SP.Rajamohana, M. Anitha, Dr. K. Umamaheswari
Assistant Professor, Department of Information Technology, Sri Shakthi Institute of Engineering and Technology, Coimbatore
Assistant Professor (Sr.Gr) Department of Information Technology, PSG College of Technology, Coimbatore-4
Assistant Professor, Department of Information Technology, Sri Shakthi Institute of Engineering and Technology, Coimbatore
Professor Department of Information Technology, PSG College of Technology, Coimbatore-4
Corresponding Author: G. Janani,

----------ABSTRACT----------
E-shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the Internet using a web browser. This popularity has made web an excellent source of gathering customer opinions about a product. Positive opinions bring significant business growth and financial gains. Similarly negative opinion cause sales loss and affect companies reputation. There is no reported study on assessing the trustworthiness of opinions, which is crucial for all opinion based applications, although web spam and email spam have been investigated extensively. Existing research is more focused towards classification and summarization of online opinions. In this work, an attempt has been made to detect whether an opinion or the review is a spam or a non-spam, to provide a trusted view to help the customer in taking a decision. The trustworthiness of the reviews is assessed as spam or a non-spam review, which includes both duplicate and near duplicate reviews classified as spam reviews, and partially related and unique reviews classified as non-spam reviews.

Keywords*: Review Spam Detection, Feature selection, Naïve Bayes, KNN Clasiifier, Classification accuracy*

Date of Submission: 28-08-2017 Date of acceptance: 04-11-2017

I. INTRODUCTION

In recent times, the contents that are available in the internet is mostly generated by the users and it is increasing rapidly. People tend to buy products are much interested in reading the reviews of the products for decision making [1]. But for these enormous reviews only little quality control prevails. Nowadays opinion spammers are hired to post spurious reviews in the website in order to promote or demote the products. The positive spam reviews about the product may lead to financial gains and helps in increasing the popularity of the product. Similarly, the negative spam reviews are posted in the intention of decreasing the fame of the product or individual. In the past few years, the problem of spam or fake reviews has been increasing. Hence, there arises a need of finding truthfulness measure of the reviews.

Feature selection (FS) is a technique which is used for selecting optimized set of features from the original set of features which may include noisy, duplicate and irrelevant features. It is mainly done to build more robust learning models and to reduce the processing cost [5]. The main goal of feature selection is to reduce the number of features thus increasing the classification accuracy and the performance. Due to the randomized nature [5], Meta-heuristics such as Particle swarm optimization (PSO) [6], evolutionary algorithms (EA) [7], Genetic Algorithm (GA), Bat Algorithm (BA) are widely used for feature selection. When the dimensionality of the feature space is high, it is quite difficult to find the optimized feature subset using the traditional optimization methods and have proven to be inefficient. Therefore, meta-heuristic algorithms are used extensively for the appropriate selection of features. Two types of feature selection methods such as the filter and wrapper approaches can be incorporated for the selection of subset of features.Wrapper approaches include a learning/classification algorithm in the evaluation procedure, while filter approaches do not. Filter approaches are argued to be computationally less expensive and more general, while wrapper approaches can usually achieve better results [15]. In this paper we proposed a Hybrid approach to select optimized feature subset from the dataset which incorporates Genetic algorithm and Bat algorithm.

The remainder of the paper is organized as follows. Section II presents the Literature Overview. Section III presents the Proposed Methodology. In Section IV we revisit the Bat Algorithm, Genetic Algorithm. In Section

V we revisit the Naïve-Bayes and KNN theory background. Section VI presents the Performance Evaluation and experimental results are discussed in Section VII. Finally, conclusions are stated in Section VIII.

II. LITERATURE OVERVIEW

Lin Shang, Zhe Zhou , Xing Liu[1] proposed a new method called F-BPSO(Fitness proportionate selection BPSO) which is used for feature selection. In this paper, to overcome the drawbacks of traditional BPSO the new method is proposed which provides a better way of calculating velocities and can solve the problem better. The traditional BPSO method, focus too much on the overall performance of a particle as a whole and thus does not pay more attention to each single dimension of it. With the analysis, it is not suitable to directly apply traditional BPSO to feature selection domain and the main reason is that the update formula of velocity is in need of modification in BPSO. There are few drawbacks in FBPSO where the first problem is that the involved particles only contains three members, namely *x, x pb* and *xgb*. Traditional PSO and BPSO only take these three particles into consideration when deciding *x*'s new velocity, but in F-BPSO a vote-like mechanism is utilized to update velocity instead of the traditional method. The voting result will be easily affected by a single voter's bias if the number of voters is too small and the second problem of original F-BPSO is the concern on fitness proportionate selection process. So this limits the sentiment classification accuracy.

Recently Yang [12] proposed a new meta-heuristic method for continuous optimization namely Bat Algorithm (BA), which is based on the fascinating capability of micro bats in to find their prey and discriminate different types of insects even in complete darkness. Such approach has demonstrated to outperform some well-known nature-inspired optimization techniques.

Susana et al.[5] have proposed an approach named modified binary particle swarm optimization (MBPSO) for feature selection and SVM for classification to mortality prediction in septic patients. They have considered 6 benchmark datasets from UCI repository to conduct experiments and to illustrate the effectiveness of the approach. MBPSO is tested in several benchmark datasets and it is compared with other PSO based algorithms and genetic algorithms (GA). MBPSO approach can correctly select the best features subset which helps to achieve high classification accuracy when compared to other PSO based algorithms. MBPSO gives less number of selected features in subset solutions when compared to Genetic algorithm.

Ahmed Majid Taha, Aida Mustapha, and Soong-Der Chen[2] proposed a new method called Naive Bayes guided Bat algorithm (BANB) which performs well in large scale data and the accuracy of the algorithm is good when compared to other algorithms. It uses the biological characteristics of a micro bat to find the optimized feature subset among the dataset. They have concluded that the proposed Naive Bayes guided Bat Algorithm (BANB) outperformed other meta heuristic algorithms with a selection of feature subsets that are significantly smaller with a less number of features. Thus the proposed algorithm reduced the number of features while at the same time increased the classification accuracy.

Iztok Fister Jr.1, Du□san Fister1, Xin-She Yang[3] In this paper the Bat algorithm is used for feature selection. Echolocation is an important feature of bat behaviour. That means, bats emit a sound pulse and listen to the echo bouncing back from obstacles whilst flying. This phenomenon has been inspired Yang [36] to develop the Bat Algorithm (BA). The algorithm obtained good results when dealing with lower-dimensional optimization problems, but may become problematic for higher-dimensional problems because it tends to converge very fast initially. In order to improve bat algorithm behaviour for higher-dimensional problems, the original bat algorithm were hybridized with differential-evolution strategies. The results of HBA is better than BA but it is not outstanding since hybrid bat algorithm is not tested on large-scale global optimization

FERHAT OZGUR CATAK, TUBITAK – BILGEM [7]. In this paper a new algorithm that considers both the number of features in feature subset and F1 score of the classifier function that is generated with this feature subset. To find the feature subset Genetic algorithm is used to obtain the optimized feature subset. F1 score is the most used model selection method in IR domain. The overall methodology is as follows: Using feature selection, the input matrix which is quite high for the memory is reduced, and matrix complexity is reduced through this way. F1 based model selection method is used. Iteration size of the algorithm is quite low. The proposed method tries to find a feature subset as small as possible while classifier hypothesis has high F1 score. Thus the proposed algorithm reduces the number of features and at the same time increases the classification accuracy. But the time taken for convergence is quite high since it involves so many parameters.

III. METHODOLOGY

The proposed methodology consists of four phases preprocessing, feature extraction, feature subset selection using Bat algorithm and Genetic algorithm and Classification using Naïve Bayes and KNN. The block diagram of the proposed methodology is illustrated in Fig. 1.

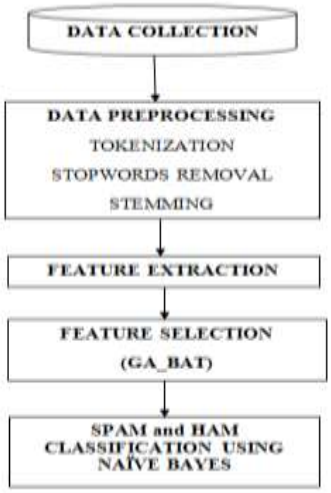

Fig 1. Block diagram of the proposed GA_BAT

A. DATA PREPROCESSING

Different pre-processing techniques were applied to remove the noise from out data set. It helped to reduce the dimension of our data set, and hence building more accurate classifier, in less time. The main steps involved are i) document pre-processing, ii) feature extraction / selection iii) SentiWordNet Score Calculation.

Data pre-processing reduces the size of the input text documents significantly. It involves activities like stop words elimination, stemming and the sentiment score detection using sentiwordnet.

- *Tokenization*

  Text document has a collection of sentences which is split up into terms or tokens by removing white spaces, commas and other symbols.
- *Stop words removal*

  Stop words are words which are filtered out prior to, or after, processing of natural language data. It removes articles like a, an, the, etc. It also removes unwanted words.

- *Stemming*

  Stemming is the action of reducing words to their root or base form. For example, using the English word "generalizations" would subsequently be stemmed as "generalizations → generalization → generalize → general → gener".

B. SENTIWORDNET

The features are extracted after preprocessing by using the SentiWordNet. The aim of SentiWordNet is to provide an extension for WordNet, such that all synsets can be associated with a value concerning the negative, positive or objective connotation. SentiWordNet 3.0 is the improved version of SentiWordNet 1.0 and publicly freely available for research purpose with a web interface. This extension labels each synset with a value for each category between 0.0 and 1.0. The sum of the three values is always 1.0, so each synset can have a nonzero value for each sentiment, because some synsets can be positive, negative or objective depending on the context in which they are used.

C. FEATURE EXTRACTION

Feature extraction helps to identify important words in a text document. This is done using methods like TF-IDF (term frequency-inverse document frequency). In the context of text classification, features or attributes usually mean significant words, multi-words or frequently occurring phrases indicative of the text category. The TFIDF relevance is calculated using equation (4).

$$\mathbf{tfidf}_{i,j} = \mathbf{tf}_{i,j} \times \log\left(\frac{\mathbf{N}}{\mathbf{df}_i}\right) \tag{1}$$

where

tf_{ij} = total number of occurrences of i in j

df_i = total number of documents containing i
N = total number of documents

IV. FEATURE SELECTION

In this section we describe the BA, BBA and GA approaches as well as the Naïve Bayes Classifier.

A. Bat Algorithm

The focus of researchers is attracted by Bat Algorithm which has the advanced capability of echolocation. Echo location works as a type of sonar, mainly micro-bats which emit a loud and short pulse of sound, and waits for it to on a object and within a fraction of time, the echo returns back to their ears of Bat [15]. Thus, bats can compute how far they are from an object [16]. In addition, this amazing orientation mechanism makes bats being able to distinguish the difference between an obstacle and a prey, allowing them to hunt even in complete darkness [17].

Based on the behaviour of the bats, Yang [12] has developed a new and interesting meta-heuristic optimization technique called Bat Algorithm. Such technique has been developed to behave as a band of bats tracking prey/foods using their capability of echolocation. In order to model this algorithm, Yang [12] has idealized some rules, as follows:

1) All bats use echolocation to sense distance, and they
Also know the difference between food/prey and background barriers in some magical way.
2) A bat bi fly randomly with velocity vi at position xi with a fixed frequency $f$min, varying wavelength λ and loudness $A0$ to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission r ε [0, 1], depending on the proximity of their target.
3) Although the loudness can vary in many ways, Yang [12] assume that the loudness varies from a large (positive) $A0$ to a minimum constant value $Amin$.

B. Binary Bat Algorithm

The concept of Binary Bat Algorithm is similar to BA Since the problem is to select or not a given feature, the bat's position is then represented by binary vectors. So the binary version of the Bat Algorithm restricts the new bat's position to only binary values using a sigmoid function. The pseudo code of BBA is shown in Algorithm 1. Initially the population is initialized randomly and the initial position xi, velocity vi and frequency fi are initialized for each bat x_i. For each time step t, being T the maximum number of iterations, the movement of the virtual bats is given by updating their velocity and position using the equations (2),(3),(4) respectively.

$$f_i = f_{\min} + (f_{\max} - f_{\min})\,\beta. \quad (2)$$

$$v_i^t = v_i^{t-1} + (x_* - x_i^t) f_i \quad (3)$$

$$x_i^t = x_i^{t-1} + v_i^t \quad (4)$$

where β denotes a randomly generated number within the interval [0, 1].

In order to improve the variability of the possible solutions, Yang [12] has proposed to employ random walks. Primarily, one solution is selected among the current best solutions, and then the random walk is applied in order to generate a new solution for each bat using equation (5).

$$x_{new} = x_{old} + \in A^t \quad (5)$$

A^t stands for the average loudness of all the bats at time t, and ϵ ranges between [−1, 1] attempts to the direction and strength of the random walk. For every iteration of the algorithm, the loudness Ai and the emission pulse rate ri are updated using equation (6) and (7) respectively.

$$A_i^{t+1} = \alpha A_i^t \quad (6)$$

$$\mathbf{ri^{t+1} = r_i^0\,[1 - exp(-\gamma t)]} \quad (7)$$

Where α and γ are constants which ranges between -1 and 1.

In BBA each solution uses a fitness function which is defined in equation (8) where $P(YJ|\,X)$ is the classification accuracy, TF is the total number of all features, and SF is the number of selected features. δ and φ are two parameters corresponding to the weight of classification accuracy and subset length, where δ ε [0,1] and $\varphi = 1 - \delta$. From (8), we can see that the importance of classification accuracy. Generally, classification accuracy is given more weight than the size of the subset. In this experiment, the two parameters have been set as follows: δ = 0.9, φ = 0.1.

$$Sol_A = \delta . P(Yj/X) + \varphi . \frac{TF - SF}{TF} \quad (8)$$

Algorithm 1: PSEUDOCODE FOR BBA_NB

```
1 begin
2 divide Dataset into a Training set and a Test set;
3 Initialize the bat population xi(i=1, 2,..., n) and   velocity vi
4 Initialize frequency fi, pulse rates ri and the loudness Ai
5 while Maximum Iterations is not reached do
6 Generate new solutions by adjusting frequency
  and updating velocities and locations/solutions
7 if(rand>ri)
        Select a solution among the best solutions
        Generate a local solution around the selected best solution
end if
8 Generate a new solution by updating the position randomly
9 if (rand <Ai& f(xi) < f(x*))
        Accept the new solutions
        Increase ri and reduce Ai
end if
10 for i = 1 to Population xi do
11 update the pbest of particle i;
12 randomly selecting a gbest for particle i from the
highest ranked bats
13 update the velocity and position of particle i
14 end
15 Rank the bats and find the current best x*
16 end while
```

C. GENETIC ALGORITHM

Genetic algorithm (GA) is an evolutionary algorithm that mimics the natural selection, crossover and mutation process. GA was first developed by Holland in 1975. GA is a stochastic optimization method, which is based on meta heuristic search procedures. GA starts with a matrix of population of solution. Each row of this matrix shows the individuals that generated randomly. Each individual shows a solution of an objective function. In GA, every solution is encoded with genes that are called individual. Using an objective function, fitness of individuals is computed according to an objective function. Population is improved with combination of genetic information from different members of population. This process is called as crossover. Another population improvement method is mutation. Some individuals of population are mutated according to the mutation rate of population.

Pseudo code of GA is show in Algorithm 2.

Algorithm 2: Pseudo code for Genetic Algorithm

```
1 begin
2 divide Dataset into a Training set and a Test set;
3 procedure GENETIC ALGORITHM(P)
4 while t=0 Initialize Population P (t) randomly
5 while Maximum Iterations is not reached do
        F(t) = Compute Fitness (P (t))
        Generate new solutions
6  t  ⟶ t + 1
7 Perform Selection
8 Perform Crosssover
        P(t) ⟶crossover(P(t - 1))
9 Perform Mutation
        P(t) ⟶mutate(P(t))
10 Compute the fitness
```

F(t) → Compute Fitness(P(t))
end while
11 return best p. » Return the best individual
12 end procedure

Initial Population: In GA, the initial populations of n
Strings are randomly generated and collection of such strings is called initial population. The solution features are represented using binary string character. Specifically 1 represents a selected attribute or feature and 0 represents the discarded one. Generate random population of n individual.

Objective Function: The objective function of the algorithm to be maximized is the sum of feature ratio and accuracy of the best chromosome.

Selection: The selection of individuals is based on the survival of the fittest. It means that bigger ones have more chance to survive and to create an offspring and to transfer the genes to the next generation. To evaluate the quality of each solution, classification accuracy is used as the fitness function. For each solution in the population, tenfold cross validation with classification algorithm is used to assess the fitness of that particular solution.

Crossover and Mutation: Crossover is the process of exchange of information between two parents to produce a new offspring. Choose two individuals from the population and perform crossover based on a crossover probability *Pc*.The probability is set to 0.6. Mutation is randomly mutated individual feature characters in a solution string based on a fixed probability *Pm*. The mutation probability is set to 0.01.
The quality of individuals is measured by fitness function. The individual is selected based on high fitness value and stored separately. The same process is repeated until the maximum iteration is reached. The genetic algorithm has fast convergence and hence it may get trapped into local optimum. So it is combined with BBA.

D) Hybrid GA_BAT
In general Genetic Algorithm has fast convergence and hence it may get trapped into local optimum. On the other hand BBA is well capable of tuning the frequency and loudness parameters to obtain the global solution. BA has a distinct advantage over other meta heuristic algorithms. That is, BA has a capability of automatically zooming into a region where promising solutions have been found. This zooming is accompanied by the automatic switch from explorative moves to local intensive exploitation. As a result, BA has a quick convergence rate, at least at early stages of the iterations, compared with other algorithms.

In the proposed work the combination of BBA and GA is done to obtain the optimized feature subset. Search space is modeled as n-cube, where n stands for the number of features. As the quality of the solution is related with the number of particles, we need to evaluate each one of them by training a classifier with the selected features encoded by the particle's quality and also to classify an evaluating set.

In the Hybrid GA_BAT the improvisation of BBA is introduced into GA to obtain the high exploration of search space so that the global optimum can be reached. Position (xi), Velocity (vi) and the Pulse rate (ri) are the key parameters of BBA which tunes to obtain the optimal solution. By combining these three parameters into GA the local best and the global best solution can be achieved efficiently. The improvement in GA is the selection of the new offspring (ie) the position of the chromosome can be updated through the BBA parameter thereby it improves the population diversity. Elitism is used to maintain the gene's quality. So the Hybrid GA_BAT fine tunes the features and it increases the classifier's accuracy.

Algorithm 3: Pseudo code for HYBRID GA_BAT

GA_PHASE
begin
1. Init population (x1,x2….xn)
2. Evaluate population
3. Repeat
   * Apply reproduction operator
   * Apply Selection operator
   * Apply crossover operator
   * Apply mutation operator

* Check for fitness
* Evaluate population

4. Until (termination condition)

BAT_PHASE

5. Select the best solution amongst the population given by last genetic pass as the initial start point
6. Start with the best solution vector and estimate the fitness (objective function)
7. Repeat
   * Generate a new solution by tuning the frequency of the bats.
   * Update the position of the bat
   * If the new position is accepted, make it the current best position, else accept the new solution as xbest
8. Do until the maximum iterations reached.

End

V. CLASSIFICATION

Classification is the process of finding a derived model which describes the data classes. The main purpose is to be able to use the model to predict the class of objects whose labels are unknown. The derived model is based on the analysis set of training data. Classification of spam reviews is not only depend upon the classifying the accurate spam review as spam but also classifying non spam reviews as normal or ham reviews**.**

A. K NEAREST NEIGHBOUR CLASSIFIER

As an initial step,1-nearest neighbor (1-nn) is employed for the classification. This 1-nn method is proposed by Duda & Hart in 1973. The implementation of this method is very easy and it does not require optimization procedure. A training sample of N vectors xj = (xj1, . . .,xjd), j = 1,...,N is assumed. In the above representation, d indicates the number of features selected and xjk is the description of observation j on feature k. In the 1 -nn classifier, an unknown observation zi = (zi1, . . .,zid) is classified based on its Euclidean distance. After calculating the Euclidean distance, zi is put into the class where its nearest training observation belongs to. The k-nn method is the extension of the 1-nn method where k-nearest neighbors are taken into consideration instead of the single neighbor as in 1-nn.

B. Naïve Bayes Classifier

Naive Bayes classifier uses the probabilistic method to predict a class for every instance of data set. It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, explicit content detection, language detection and sentiment detection. Even though it is often outperformed by other techniques such as boosted trees, random forests, Max Entropy, Support Vector Machines etc, Naive Bayes classifier is very efficient since it is less computationally intensive (in both CPU and memory) and it requires a small amount of training data. Moreover, the training time with Naive Bayes is significantly smaller as opposed to alternative methods. Naive Bayes classifier is superior in terms of CPU and memory consumption and in several cases its performance is very close to more complicated and slower techniques. The specific working process of Naive bayes is as follows:

Let T be the training sample set. Each sample has category labels. Sample set has a total of m classes: C1, C2,...Cm. Each sample is represented by an n-dimensional vector System design X={x1, x2, ..., xn}, and each vector describes n attributes A1, A2, ..., An. The different ways in calculating the probability of the class is explained below.

1. Given a simple X, the classifier will predict that X belongs to the highest posterior probability of class. If and only if $P(C_i|X)>P(C_j|X)$, $1<=i, j<=m$, X is predicted to belong to class C_i. According to the bayes' theorem, the probability is calculated as in equation (9).

$$P(C\_i/X)=(P(X/C\_i)*P(C\_i))/(P(X) \quad (9)$$

Because P(X) is the same for all classes, it only need to find the largest $P(X|C_i)P(C_i)$. The prior probability of class C_i can be calculated. $P(C_i)=s_i/s$, s_i is the number of training samples of class C_i, and s is the total number of training samples. If the prior probability of class C_i is unknown, it is usually assumed that the probability of these classes are equal, then P(C1)= P(C2)= …= P(Cm), therefore the problem is transformed into how to get maximum $P(X| C_i)$.

2. If the data set has many attributes, the workload of calculating $P(X| C_i)$ is very high. In order to reduce the computational overhead of $P(X|C_i)$, simple assumptions that under certain condition attribute characteristic value is independent of each other. $P(X| C_i)$ is calculated as in equation (3)

$$P\left(\frac{X}{C_i}\right) = \prod_{k=1}^{n} P\left(\frac{x_k}{C_i}\right) \quad (10)$$

3. Probability $P(x_1|C_i), P(x_2|C_i), \ldots, P(x_n|C_i)$ can be calculated from the training set. Here x_k refer to the attribute A_k of sample X.

4. For each class, calculating $P(X|C_i)P(C_i)$. If and only if $P(X|C_i)P(C_i)$ is maximum, the classifier prediction sample X belongs to class C_i. We have used the bayes' theorem for classification as the past information about a parameter can be incorporated and form a prior distribution for future analysis.

VI. PERFORMANCE EVALUATION

We evaluate the classification performance in terms of three commonly used metrics: accuracy, recall and precision as defined in equation 11-13 and Table 1. Table 1 is a confusion matrix whose entries are given as a function of two typical classes in review spam detection, positive and negative reviews.

$$Accuracy = \frac{\text{True positive} + \text{True Negative}}{\text{True positive} + \text{True Negative} + \text{False Negative} + \text{False Positive}} \quad (11)$$

$$\text{Precision} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{False positives}} \quad (12)$$

$$\text{Recall} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{False negatives}} \quad (13)$$

VII. EXPERIMENTAL STUDY

The Proposed GA_BAT algorithm was implemented using Net beans IDE 8.0 of the system with Intel P4, 2.66GHz CPU; 16 GB RAM with Windows XP Professional. In our experiments, GA_BAT feature selection algorithm is used to select the optimized feature subsets from the review spam dataset. The stages of the proposed methods results are presented below.

A) Dataset description

The dataset used for the proposed method was assembled by Ott et al. (2011) and Ott et al.(2013). The dataset consists of 1600 reviews of the 20 most popular Chicago hotels which is organized as follows: 800 positive reviews of which 400 are truthful and 400 are deceptive, 800 negative reviews of which 400 are truthful and 400 are deceptive [27]. From this review dataset, 80% (1280 instances) of the reviews are taken for the training process and the other 20% (320 instances) of reviews are taken for the testing process with the significant features.

B) Parameter settings for BBA:

The parameters used for the proposed BBA and GA are shown in the Table 2 and Table 3. After initializing the parameters, the fitness function is calculated using classification performance. It is used to evaluate the selected subsets of features for each dataset. The training process was implemented using 10 fold cross validation method.

Table 1: Parameter settings for BBA

Population Size	30
Loudness	0.8
Pulse rate	0.2
Frequency range	20 and 500 KHz

Table 2: Confusion Matrix

C) Parameter settings for GA:

Population size	30
Crossover rate	0.6
Mutation rate	0.05

Table 3: Parameter settings for GA

	True Positive reviews	True Negative Reviews
Actual Positive Reviews	TP	TN
Actual Negative Reviews	FP	FN

From this review dataset, 80% (1280 instances) of the reviews are taken for the training process and the other 20% (320 instances) of reviews are taken for the testing process with the significant features. The average length of a single review is around 100 characters. Example reviews of datasets are given in Table 1 and our target is to classify such reviews into two categories: truthful reviews and deceptive reviews.
Before running feature selection and classification method, the review spam dataset is initially preprocessed. After tokenizing a document, commonly used terms or stop words are first removed from the term set of each document. The number of features after stemming is decreased when compared to the features after stop words removal. The number of features obtained in every step is shown in the table 3.

Total no. of Reviews	**1600**
No. of Terms	239725
After Stop Words Removal	118735
After Stemming	117803
Senti wordnet Features With Duplicates	53648
Sentiwordnet Features Without Duplicates	1771

Table 4: Feature count comparison for preprocessing

We apply the well known sentiment lexicon called SentiWordNet in the proposed scheme which contains about 10000 words and their sentiment polarities either as positive or negative. Similarly, after SentiWordNet the feature count has been reduced to 53648. Moreover the features after SentiWordNet have many duplicate features. These redundant features are removed. Thereby, the feature count has been reduced to 1771. Table 4 shows the sentiment scores and TFIDF values for features.

Review no	**Features**	**Sentiment Score**	**TFIDF values**
11	Nice	0.06483843537414967	0.002822
15	Old	-.00984210449089646	0.005230
8	Good	0.05113636363636363	0.003268
13	Room	-0.3926601666079218	0.001249
4	Cheap	0.3584408049559635	0.007427

Table 5 Sentiment Scores and TF-IDF Values for Sample Features

After generating the TF-IDF matrix, the parameters of BBA are initialized. These TF-IDF values are given as an input to the BBA and GA. In BBA, the position values are randomly initialized with 0 and 1.The fitness value for each features are calculated using the TF-IDF values of the selected feature. The local best value and the global best values are initially noted. The frequency, loudness, velocity and the position values are updated till the maximum iteration is reached.
In Genetic Algorithm, Roulette wheel selection is applied to select the best parents, reproduce offspring by random single point crossover and then perform flip bit mutation for optimized feature selection. The proposed GA_BAT evolutionary operators are modified that improves performance. An optimized feature is selected by GA_BAT that is classified using Naïve Bayes. This kind of combination significantly reduces computational cost and also increases classification rate. The genetic algorithm GA_BAT based Naïve Bayes performance is better than KNN classification.
The figure 2 depicts the comparison of different feature selection techniques. In which GA_BAT with NB achieves the greater performance than the other feature selection techniques.

Correctly Classified Instances	Incorrectly Classified Instances	Mean absolute error	Root mean squared error	Relative absolute error	Coverage of cases (0.95 level)	Mean rel. region size (0.95 level)
45	35	0.4709	0.6533	111.4%	61.25 %	63.125%

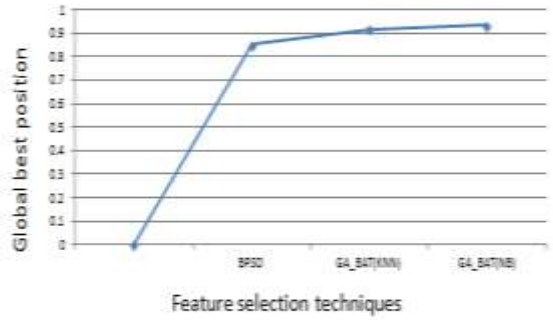

Fig 2: Performance analysis of feature selection methods

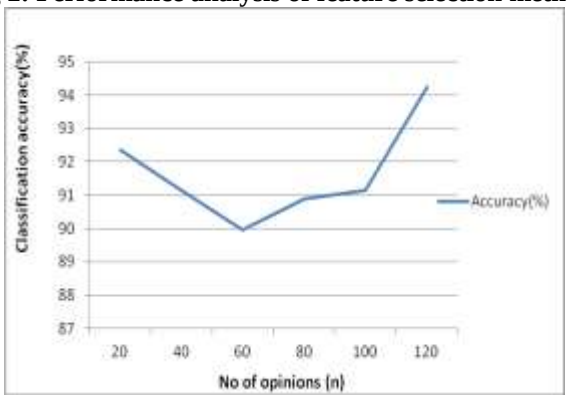

Fig 3: Classification Accuracy

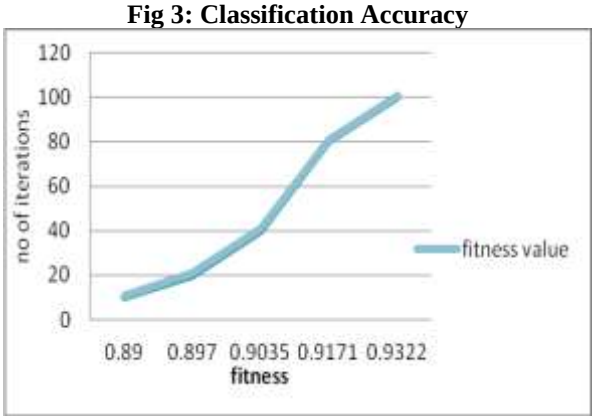

Fig 4: Comparison of fitness across various iterations

No of Reviews	20	40	60	120
Accuracy (%)	90.35	94.14	89.14	91.45

Table 6 Classification Accuracy

Table 7: Classifier output analysis

No. of reviews (n)	Time complexity (ms)
	GA_BAT-NB
20	40.88
40	52.05
60	59.32
80	65.49
100	73.62
120	79.75
140	90.83

Table 8: Time Complexity analysis

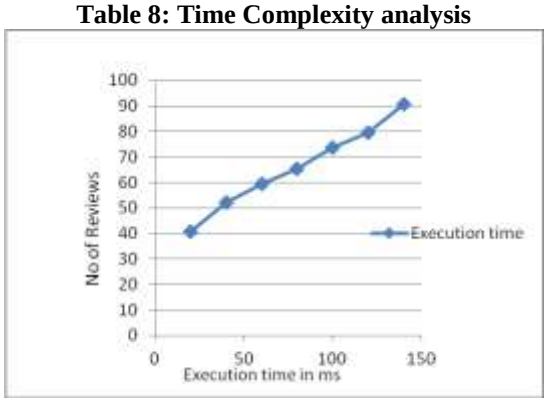

Fig 5: Comparison of Time Complexity

VIII. CONCLUSION

In this paper, we have proposed a hybrid method which integrates Binary Bat algorithm and the Genetic algorithm for feature selection using Naïve Bayes classifier for review spam detection. An efficient review spam classification using hybridized GA_BAT reduces the computational time for obtaining review spam by applying Naïve Bayes classifier with the optimal features selected. This in turn improves the classification accuracy of the online review spam that efficiently classifies the fake and real review at lesser time interval.

REFERENCES

[1]. Zhe Zhou, Xing Liu, Ping Li, and Lin Shang "Feature Selection Method with Proportionate Fitness Based Binary Particle Swarm Optimization" 10th International Conference, Proceedings pp 582-592, December 15-18, 2014.
[2]. AhmedMajid Taha, Aida Mustapha, and Soong-Der Chen "Naive Bayes-Guided Bat Algorithm for Feature Selection" The Scientific World Journal Volume (2013)
[3]. Iztok Fister Jr, Du˘san Fister, Xin-She Yang "A Hybrid Bat Algorithm" Faculty of electrical engineering and computer science, Smetanova 17, 2000.
[4]. R. Nakamura, L. Pereira, K. Costa, D. Rodrigues, J. Papa, and X. S. Yang, "BBA: a binary bat algorithm for feature selection," in Proceedings of the 25th Conference on Graphics, Patterns and Images (SIBGRAPI '12), pp. 291–297, 2012.
[5]. Susana M. Vieira, Luís F. Mendonça, Gonçalo J. Farinha and João M.C. Sousa, " Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients", Applied Soft Computing, vol 13,pp.3494-3504-248, 2013.
[6]. J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825–1844, 2007.
[7]. A. M. Taha and A. Y. C. Tang, "Bat algorithm for rough set attribute reduction," Journal of Theoretical and Applied Information Technology, vol. 51, no. 1, 2013.
[8]. X.-S. Yang., "Bat algorithm for multi-objective optimisation," International Journal of Bio-Inspired Computation, vol. 3, no. 5, pp. 267–274, 2011.
[9]. A. H. Gandomi, X. S. Yang, A. H. Alavi, and S. Talatahari, "Bat algorithm for constrained optimization tasks," Neural Computing and Applications, vol. 22, pp. 1239–1255, 2012.
[10]. A. M. Taha and A. Y. C. Tang, "Bat algorithm for rough set attribute reduction," Journal ofTheoretical and Applied Information Technology, vol. 51, no. 1, 2013.
[11]. R. Nakamura, L. Pereira, K. Costa, D. Rodrigues, J. Papa, and X. S. Yang, "BBA: a binary bat algorithm for feature selection," in Proceedings of the 25th Conference on Graphics, Patterns and Images (SIBGRAPI '12), pp. 291–297, 2012.
[12]. Zhipeng Liu, Dechang Pi, and Yunfang Chen, "Mining Potential Spammers from Mobile Call Logs", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Volume 2015, , Pages 1-9, September 2014.
[13]. Y. H. Li and A. K. Jain, "Classification of Text Documents", the computer journal, vol. 41, no. 8, 1998.
[14]. Pravesh Kumar Singh, Mohd Shahid Husain, "Analytical Study of Feature Extraction Techniques In Opinion Mining", Department of Computer Science and Engineering, 2013

[15]. M. Anbarasi, E. Anupriya, and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," International Journal of Engineering Science and Technology, vol. 2, pp. 5370–5376, 2010.
[16]. S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," Expert Syst. Appl., vol. 35, no. 4, pp. 1817–1824, 2008.
[17]. L. Y. Chuang, S. W. Tsai, and C. H. Yang, "Improved binary particle swarm optimization using catfish effect for feature selection," Expert Syst. Appl., vol. 38, no. 10, pp. 12 699–12 707, 2011.
[18]. Omar S.Soliman, Eman Abo Elhamd, "Classification of Hepatitis C Virus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine", International Journal of Scientific & Engineering Research, ISSN 2229-5518, Volume 5, Issue 3, 2014.
[19]. Mita K. Dalal, Mukesh A. Zaveri, "Automatic text classification: a technical review", International Journal of Computer Applications (0975 – 8887)Volume 28– No.2, August 2011.
[20]. Y. H. Li and A. K. Jain, "Classification of Text Documents", the computer journal, vol. 41, no. 8, 1998
[21]. Pravesh Kumar Singh, Mohd Shahid Husain, "Analytical Study Of Feature Extraction Techniques In Opinion Mining", Department of Computer Science and Engineering, 2013.
[22]. S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," Expert Syst. Appl., vol. 35, no. 4, pp. 1817–1824, 2008.
[23]. L. Y. Chuang, S. W. Tsai, and C. H. Yang, "Improved binary particle swarm optimization using catfish effect for feature selection," Expert Syst. Appl., vol. 38, no. 10, pp. 12 699–12 707, 2011.
[24]. Omar S.Soliman, Eman Abo Elhamd, "Classification of Hepatitis C Virus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine", International Journal of Scientific & Engineering Research, ISSN 2229-5518, Volume 5, Issue 3, 2014.
[25]. I. Ahmad and F. Amin, "Towards feature subset selection in intrusion detection," in Proceedings of the IEEE 7th Joint International Information Technology and Artificial Intelligence Conference (ITAIC '14), pp. 68–73, Chongqing, China, 2014.
[26]. Abdul-Rahman, A.A. Bakar, Z.A. Mohamed-Hussein, Optimizing big data in bioinformatics with swarm algorithms., In: 16th International Conference on Computational Science and Engineering (CSE), Sydney, NSW, IEEE, pp.1091–1095, 2013.
[27]. Abd.Samad Hasan Basari,Burairah Hussin and et.al,"Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization",Elsevier,2012.
[28]. KusumKumari Bharti, Pramod Kumar Singh, "Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering", Appl. Soft Computing, vol. 43, pp.20-34, 2016.

G. Janani. "Feature Selection Using Hybrid Approach for Opinion Spam Detection." The International Journal of Engineering and Science (IJES), vol. 6, no. 9, 2017, pp. 61–72.