# Feature Selection for Opinion Mining using Binary Cuckoo Search Algorithm

**S. M. Hemalatha[1] C. S. Kanimozhi Selvi[2]**
[1]PG Scholar [2]Professor
[1,2]Department of Computer Science & Engineering
[1,2]Kongu Engineering College, Perundurai, Erode, India

*Abstract*— Feature selection is a process of identifying the related subset with less intricacy. It is used to handle the optimization problem in classification. The dataset may comprise enormous number of features so it is difficult to identify the related features because the dataset may contain unrelated features. To overcome this problem, a binary cuckoo search based feature selection algorithm (BCSA) is proposed. It enhances the process of feature selection and produces the best optimal feature subset which increases the predictive accuracy of the classifier with minimum number of features. It improves the search space both global and local. In this BCSA is used as a feature selector and yields the feature subset and SVM classifier is used to estimate the feature subset produced. The analysis result show that the SVM classification yield enhanced accuracy when binary cuckoo search algorithm is used for feature selection.

***Key words:*** Feature Selection, Opinion Mining, Binary Cuckoo Search Algorithm

## I. INTRODUCTION

Data mining is the process of converting into meaningful format from a large data set the main aim of this data mining is to identify the model It also predict future trends[17][16][14]

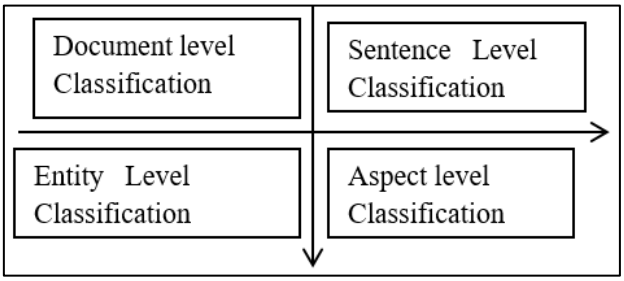
Sentimental analysis is the area of study that examines people opinions, expressions, evaluations, actions and character towards entities such as products, services, organizations, individuals, concerns, measures, focuses and their aspects[12][11][14]. It is also called opinion mining. The sentiment analysis task is subjectivity investigation, sentimentality ordering, opinion junk recognition; opinions are essentials to all human activities because they are the ideal influencers of our actions. Whenever we required making a resolution, we want to know others opinions. In the real world, businesses and establishments always want to find customer or unrestricted opinions about their products and services. Individual customer also want to know the thoughts of pre workers of a product before obtaining it and others thoughts about political candidates before making a voting decision in a political election[12][11][14].

In the past, when a specifically need opinions, he/she approached friends and family. When an association or a business needed public or consumer opinions, it accompanied surveys, opinion polls and motivates the groups. Attaining public and consumer opinions has long been a huge business itself for marketing, public relations and political campaign companies. Analyzing such composed features and mining specific model from the text is in main focus that leads to the growth of sentiment analysis [12][11][14]

### *A. Levels of Sentiment Classification*

The sentiment classification has been divided mainly into three levels:

- Document level classification
- Sentence level Classification
- Word Level Classification




The fig 1.1 represents the level of sentimental Classification
The main task of this level is to separate whether an entire opinion document expresses a positive or negative sentiment. For example, a product reviews the system that defines whether the review states a globally positive or negative opinion about the product. It is also known as document level sentiment classification. This level of analysis assumes that each document expresses thoughts on a single entity. Thus, it is not suitable to all documents which estimate or compare manifold entities [12][11].

Both the document level and the sentence level classification do not determine what exactly people liked and did not like. Aspect level performs finer grained analysis. Instead of looking at language constructs such as documents, paragraphs, sentences, clauses or phrases, aspect level directly looks at the opinion itself. An opinion without its goal being recognized is of limited use realizing the significance of opinion target makes to comprehend the sentiment analysis problem better [12][11].

### *B. Sentiment Classification at Document Level*

Document level is an identification method that classifies the opinion of the whole document as positive, negative or neutral .It is done based on the sentence level and word level sentiment classification value[12][11]

### *C. Sentence Level Sentiment Analysis*

Sentence level sentiment analysis classifies the opinion of a sentence and accomplishes it as positive, negative or neutral. Sentence polarity classification in each text file or documents is based on word level sentiment analysis [12][11].

### *D. Word Level Sentiment Analysis*

Word level sentiment analysis identifies the opinion of a word and accomplishes as positive, negative or neutral. Word polarity classification in every text file or documents is based on a word. It mines the object attributes [12][11]

### *E. Feature for Sentimental Analysis*

Feature selection is method of selecting a subset of the terms arising in the trained data's are used only subset in text

classification. Feature selection serves two main purposes. First, it makes trained and applies a classifier more efficient by decreasing the size of the text document. Second, feature selection is often to increases classification accuracy. Feature selection are used to the execute document classification, increases the classification performance and reduce the complexity of extracted data. It should be differentiated from feature extraction. The feature extraction method creates original features that feature is based on functions and that functions are based on unique features, whereas feature selection returns a subset of the features. Feature selection techniques are mainly used in domains datasets [12][11][14].

## II. LITERATURE REVIEW

Hemalatha, kanimozhi Selvi (2018) Introduced a feature Selection for opinion mining using Shuffled Frog Leaping Algorithm In this work, biomedical opinions are extracted from twitter which contains many features needed to classify the opinions. However, such datasets contain many irrelevant or weak correlation features which influence the predictive accuracy of classification. SFLA is implemented as a feature selector and the feature subset is generated and SVM classifier is used to evaluate the feature subset produced.SFLA algorithm optimizes the process of feature selection and yields the best optimal feature subset which increases the predictive accuracy of the classifier. Experimental results show that the SVM classification produces an increase in accuracy of 7% when the selected features from shuffled frog leaping algorithm are used[14].

Rodrigues, et al (2013) introduced a binary cuckoo search algorithm for feature selection for find the most discriminative set of features can enhance the recognition rates and also to make feature extraction faster. The main objective of this work introduced a new feature selection called Binary Cuckoo Search, which is based on the performance evaluation of cuckoo birds. The researches were carried out in the context of theft recognition in power distribution systems in two datasets obtained from a Brazilian electrical power company, and have confirmed the strengthens of the proposed method against with numerous nature-inspired optimization methods[6].

Senthil nath et al (2013) proposed a Clustering using Levy Flight Cuckoo Search with three nature-inspired algorithms namely Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Cuckoo Search (CS) on clustering problem. Cuckoo search is used with levy flight. The heavy-tail property of levy flight is oppressed here. These algorithms are used on three standard benchmark datasets and one real-time multi-spectral satellite dataset. The results are calculated and examined using various techniques. Finally we accomplish that under the given set of parameters, cuckoo search works efficiently for majority of the dataset and levy flight plays a significant role[7].

## III. APPROACHES OF FEATURE SELECTION

### *A. Classes of Search Technique*

In optimization of a design, the main advantage and disadvantage of this search technique simply to minimize the cost of production or to maximize the efficiency of production. An optimization algorithm is a technique which is accomplished iteratively by comparing various resolutions till an ideal or satisfactory resolution is found [5][10][13][14][2].

A good proportion of this search time will be spent optimizing the components placement in the layout. In searching for optimum solutions, optimization techniques are used and can be divided into three broad classes [10][5][2]

*1) Numerical Technique*

This technique uses a set of necessary and suitable conditions to be fulfilled by the solutions of an optimization problem. They are segmented into direct and indirect methods. Indirect methods search for local extremes by solving the usually non-linear set of equations resulting from setting the gradient of the objective function to zero. The search for probable solutions (function peaks) starts by limiting itself to points with zero gradient in all directions. Direct methods, such as those of Newton or Fibonacci, seek excess by "hopping" around the examine space and evaluating the gradient of the new point, which guides the search [2][5].

*2) Enumerative Technique*

Enumerative technique search each point related to the function's domain space i.e. one point at a time. The advantage is easy to implement, disadvantage is computationally intensive and these techniques are not suitable for applications with large domain spaces. Dynamic programming is another example of these techniques [5][13][10][2].

*3) Guided Random Search Technique*

The guide random search technique is based on enumerative techniques with additional information to guide the search. Two subclasses are simulated annealing and evolutionary algorithms. Both can be seen as evolutionary technique but simulated annealing mainly used in a thermodynamic evolution process with minimum energy states. Evolutionary algorithms use natural selection principles. This form of search changes throughout generations refining the features of potential solutions by means of natural inspired algorithm. SFLA are best example of this technique [14][2].

### *B. Binary Cuckoo search Algorithm*

Binary cuckoo search is an extension of cuckoo search is an optimization algorithm It was stimulated by the obligate brood parasitism of some cuckoo species by hatching their eggs in the shells of other swarm birds (of other species). Some swarm birds can involve direct fight with the intruding cuckoos [3][4][6][7][8][9][1]. For example, if a swarm bird regulates the eggs are not their individual, it will either chuck these alien eggs away or destroy its nest and construct a new nest elsewhere. Some cuckoo kinds such as the New World brood-parasitic [4][3][6]

### *C. Binary Cuckoo search Representations*

Each egg in a nest denotes a solution, and a cuckoo egg denotes a new solution. The main objective of this algorithm is to use the original and potentially well enhanced solutions (cuckoos) to interchange a not-so-good solution in the nests [7][4][1]. In the easiest form, each nest has one egg. The algorithm can be protracted to more complex cases in which each nest has numerous eggs representing a set of solutions[6][3][4][9][8][1].

*1) Binary Cuckoo Search Rules*

- Each cuckoo lays one egg at a time, and scrapyards its egg in a arbitrarily selected shell[3][9][10][4][1];
- The best nests with high feature of eggs will convey over to the next generation[7][8][9][4][1];
- The number of available swarm nests is static, and the egg laid by a cuckoo is exposed by the swarm bird with a probability Determining operate on some set of worst nests, and determinate solutions deserted from beyond calculations[7][3][4][6][1].

*2) Levy Flight*

Levy flight is a random walk between cuckoo egg and cuckoo nest. This random walk can be observed in animals and insects. (1.1) represent the levy flight equation with random walk [7][3][4][6][9][8][1]

$$x_i^j(t) = x_i^j(t-1) + \alpha \oplus \text{levy}(\lambda) \longrightarrow (1.1)$$

## IV. WORKING STEPS OF BCSA

Structure Architecture

The main working phases of the proposed feature selection method are represents as given in Figure 4.1. Each phase is defined as follows:

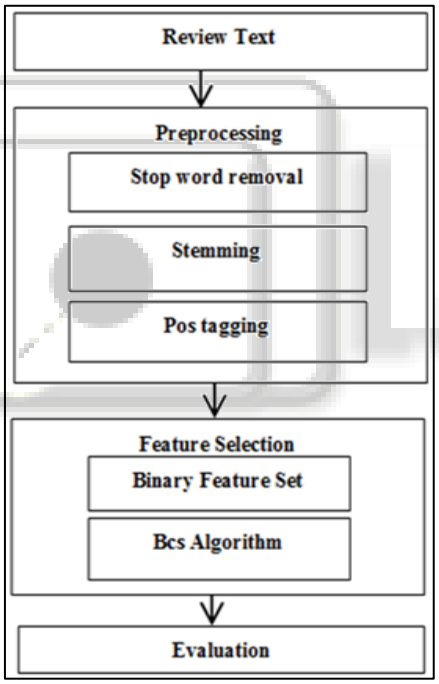



Fig 4.1 Structure Architecture of BCSA

### A. Proposed System

The proposed system collects number of positive and negative opinion. The elements of the proposed system are the modules of the proposed system

1) Data Preprocessing
2) Feature selection
3) Classification

### B. Data Preprocessing

It is a data mining method that converts raw data into an understandable format. Practical data is often imperfect, unreliable, and deficient in certain performances or trends, and is likely to contain various

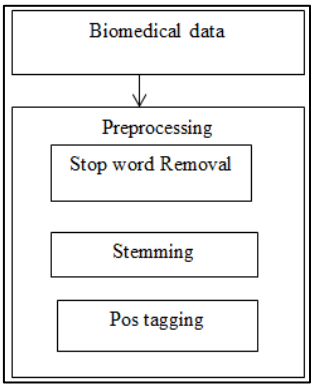



Fig. 4.2: Flow Diagram of Data Preprocessing

It is used to remove unrelated data in reviews.

*1) Stop Word Removal*

Most recurrently used words in English are not suitable in text mining. Such words are called stop words. It is a language explicit functional words which convey no data. It contain of pronouns, prepositions, conjunctions. Stop word removal is used to remove irreverent words in each review sentence. Words like is, are, was etc. Reviews are stored in text file that text file is given as input to stop word removal. Stop word is uninvolved by checking against stop words list.

*2) Stemming*

Stemming is used to form root word. It reduces the words "facing", "faced", and "facer" to the root word, "face". In this stemming contain many algorithms like n-gram analysis, Affix stemmers and Lemmatization algorithms. Porter stemmer algorithm is used to form root word for a given particular input reviews and that reviews store it in a text file.

*3) POS Tagging*

The Part-Of-Speech is a verbal category that is defined by its linguistic or semantic performance POS tagging is the function of classifying each word in a sentence with its applicable part of speech. POS tagging is a significant segment of opinion mining, it is essential to consider the features and opinion words from the reviews. It contain two phase to perform the pos tagging first phase can be done by manually and second phase can be done with help of POS tagger. Manual POS tagging of the reviews take more time to complete its process. Here, POS tagger is used to label all the words of reviews. Stanford tagger is used to label each word in a review sentence. Finally nouns are collected and stored in a text file.

### C. Feature Selection

The features are selected according to the relevance of the feature algorithm of BCS based feature selection is given below:

*1) Steps of BCSA based Feature Selection*

1) Step1: Extract Bio medical Data from Twitter
2) Step 2: Apply Preprocessing
   Stop Word Removal
   Stemming
   Pos Tagging
3) Step3: Extract Noun Using Pos Tagging
4) Step 4: Initialize the parameter (N=5, T=10)
5) Step 5: Create binary set of feature with class label
6) Step 6:For each nest=1to N
7) Step 7:For each Iteration=1to T;

Fi=-INF; end
Global fit=-INF ; end

8) Step 8: Find the Global best , Local worst value from the nest
9) Step 9: Create training and testing set using svm classifier
10) Step10: Find the accuracy for each nest and store the accuracy in acc
11) Step 11:If acc<Fi
12) then substitute acc=Fi
13) Step 12:Maxfit,Maxindex=Max(f)
14) Step 13:If Maxfit >Globalfit
15) Step 14:Globalfit=Maxfit
16) Step 15:Update Local worst using levy flight equation(1)
17) Step 16:Finally Return the optimal global best values

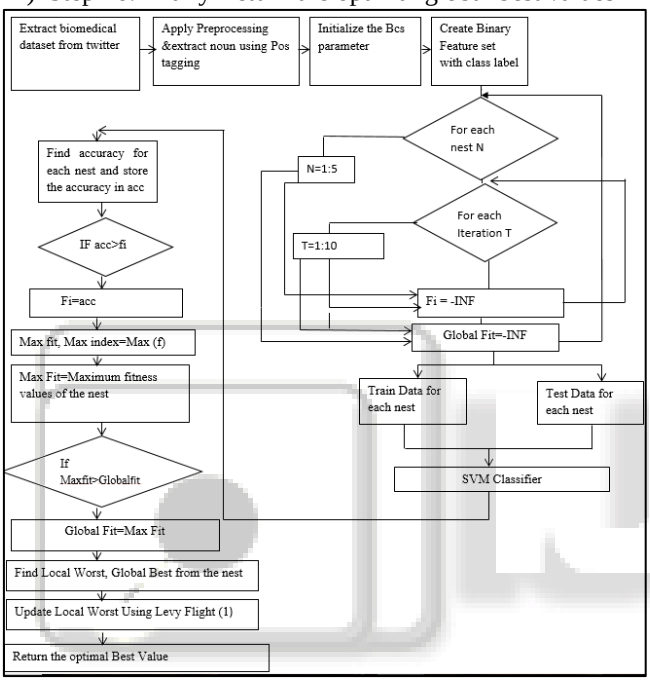



Fig. 4. 3: Feature Selection Process with BCS Algorithm

### D. *Classification*

Support Vector Machines (SVM) is a method of classification techniques the main objectives of this algorithm is divides the data using hyper planes. It is used to create multiple extrication hyper planes [16]

In this work, the Binary Feature Set with class label is given as input to the BCS algorithm and BCS is implemented as a feature selector and the binary feature subset is generated and SVM classifier is used to evaluate the Binary feature subset with train and test data.

## V. RESULT CONSVERATON OF BCSA

### A. *Result of BCSA*

Twitter account has been generated and 1500 tweets were collected with biomedical breast cancer". In the preprocessing steps, retweets were removed and only 150 tweets are reflected for advance dispensation.

#### 1) *Performance Estimation*

Tweets are classified typically as positive and negative tweets. 80% of the tweets are given for training and 20% are given for testing. SVM classification is performed in two phases.

In the first phase, the original tweets after pre-processing that convert into binary feature set with selected features are given as input to the svm classifier.

In the second phase, the Construct train and test data for each nest set with binary feature set of selected features and store in a matrix file that matrix file is passed to binary cuckoo search algorithm that input evaluate it using svm classifier and find the accuracy for each nest.

### B. *Evaluation of the Results*

#### 1) *Precision*

It is a proportion of Predicted Positive cases that are correctly Real Positives[15].

#### 2) *Recall*

Recall (also known as sensitivity) is the proportion of Real Positive cases that are correctly Predicted Positive [15].

#### 3) *F-Measure*

It is a combination of both precision and recall is i.e. harmonic mean of precision and recall, the traditional F-measure or balanced F-score[15].TABLE I represents the Evaluation of the natural inspired algorithm

| Algorithm | BCS | SFLA |
|---|---|---|
| PRECISION | 0.88 | 0.82 |
| RECALL | 0.85 | 0.79 |
| F-MEASURE | 0.432 | 0.42 |
| ACCURACY | 0.73 | 0.91 |

Table 1: Evaluation of Algorithm

#### 4) *Precision*

The TABLE II illustrate that the precision values in SFL and BCS algorithm

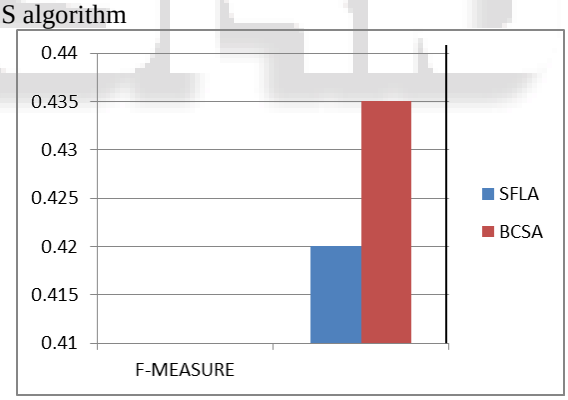



Table 2: Precision Values in SFL and BCS Algorithm

#### 5) *Recall*

The TABLE III illustrate that the recall values in BCS and SFL Algorithm

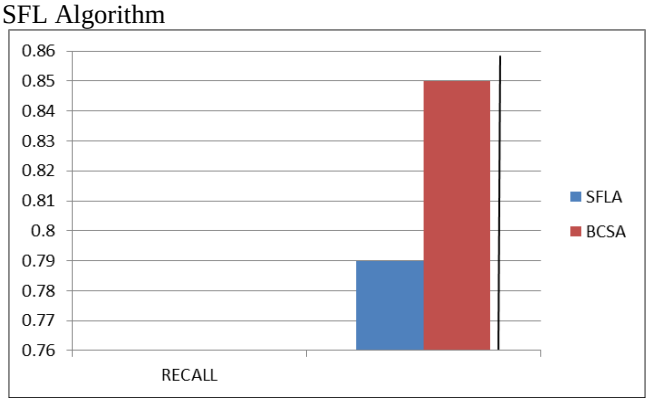



Table 3: Recall Values in BCS and SFL Algorithm

*6) F-Measure:*

The TABLE IV illustrate that the F-Measure Values in SFL and BCS Algorithm

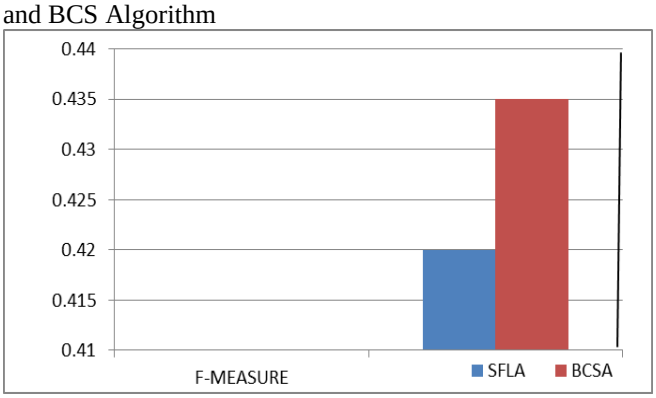



Table 4: F-Measures Values in SFL and BCS Algorithm

*7) Accuracy*

The TABLE V illustrate that the Accuracy Values in SFL and BCS Algorithm

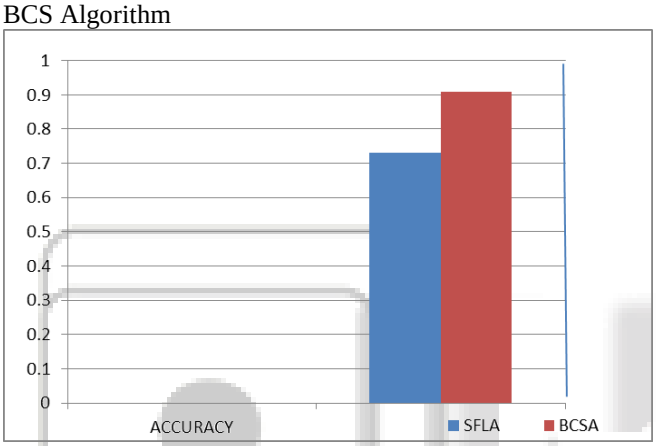



Table 5: Accuracy Values in SFL and BCS Algorithm

The accuracy of BCS algorithm 18% improves its accuracy when compared to SFL algorithm

## VI. CONCLUSION OF BCSA

The main objective of this feature selection is to not only classify a feature subset from unique set of features but also to reduce the intricacy in data mining. Without a feature selection algorithm, it is difficult to identify the pattern for the existing classification techniques in this phase, biomedical data are collected from twitter account. It contains various features desired to classify the opinions. However, such datasets contain many unrelated or weak association features which impact the predictive accuracy of classification. In this work a binary cuckoo search based feature selection algorithm (BCSA) is proposed which enhances the process of feature selection and return the best ideal feature subset which increases the predictive accuracy of the classifier with minimum number of features

BCS algorithm enhances the process of feature selection and yields the best optimal value with minimal number of feature which increases the predictive accuracy of the classifier when related to SFL Algorithm in this wok analysis result show that the SVM classification yield better accuracy when binary cuckoo search algorithm is used for feature selection.

## REFERENCES

[1] Gandomi, A., Yang, X.S., Alavi, A: Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. Engineering with Computers 29(1), 17–35 (2013)
[2] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn.Res. 3, 1157–1182 (2003)
[3] Kaveh, A., Bakhshpoori, T.: Optimum design of steel frames using cuckoo search algorithmwith levy flights. The Structural Design of Tall and Special Buildings pp. n/a–n/a (2011)
[4] Pereira L.A.M,Rodriguel.D, Papa J.P, Bianry Cuckoo Search and its Application for Feature Selection,Research Gate (2014)
[5] Papa, J., Pagnin, A., Schellini, S., Spadotto, A., Guido, R., Ponti, M., Chiachia, G., Falcao A.: Feature selection through gravitational search algorithm. In: Proceedings of the 36th IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2052–2055 (2011)
[6] Rodrigues, D., Pereira, L.A.M., Almeida, T.N.S., Ramos, : BCS: A binary cuckoo search algorithm for feature selection. In: Proceedings of the IEEE International Symposium on Circuits and Systems. Beijing, China (2013)
[7] Senthilnath, J., Das, V., Omkar, S., Mani, V.: Clustering using levy flight cuckoo search. In:J.C. Bansal, P. Singh, K. Deep, M. Pant, A. Nagar (eds.) Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), Advances in Intelligent Systems and Computing, vol. 202, pp. 65–75. Springer India (2013)
[8] Venkata,Vijaya Geetha, Ravi Krishna Kumar: Cuckoo search optimization and its application proceedings of International Journal of research in Advanced computer and communication Engineering.
[9] Mana sopa,Niwat Angka wisittpan: An Application of Cuckoo Search algorithm for series system with cost and multiple choice constraints proceedings of International Electrical Engineering Congress.
[10] Hu, Bin,. "Feature selection for optimized high-dimensional biomedical data using the improved shuffled frog leaping algorithm." IEEE/ACM transactions on computational biology and bioinformatics (2016).
[11] Liu, Bing. "Sentiment analysis and Opinion mining." Synthesis lectures on human language technologies 5.1, 1-167, and 2012.
[12] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2.1–2, 1-135, 2008.
[13] Jorge Vergara. A.Pablo Estevez, "A Review of feature selection methods based on mutual information," Neural Computation & Application, vol.24, pp. 175-186, 2014.
[14] Hemalatha , Kanimozhi Selvi "Feature Selection for Opinion Mining Using Shuffled Frog Leaping Algorithm".The International Journal of Engineering and Computer Science ,Volume 7 Issue 2, 2018
[15] powers, "Evaluation from Precision, Recall and F-Measure to roc, Informedness, Markedness &

Corrleation" Journal of Machine Learning Technologies,2011

[16] C. Kalaichelvi, and K. Selvi. "Frequent itemsets generation using collective support threshold for associative classification." National Conference on Recent Trends in Communication and Signal Processing. Vol. 2009.

[17] C.S .Kanimozhiselvi, and A. Tamilarasi. "Mining of High Confidence Rare Association Rules with Automated Support Thresholds." European Journal of Scientific Research 52.2, 2011.

