International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

# Building Large Scale Cloud System for Product Sentiment Analysis using Genetic Algorithm Based Feature Selection

**P Vasudevan, K P Kaliyamurthie**

***Abstract*: *In Sentiment analysis, any data driven approach involves changing a piece of text into a feature vector. An optimization scheme of the best-first search which decreases the amount of memory required is referred to as beam search. The possibility of the Beam Search finding the goal can be improvised using a more precise heuristic function as well as a greater beam width. This work covers the local beam search based on feature selection and Genetic Algorithm (GA). A subset of features can be found utilizing the GA where, the bits of chromosomes indicate the presence or the absence of features. Also, for obtaining the best sub-optimal set, the global maximum for the objective function can be discovered. Here, the performance of the predictor is the objective function. As the performance of Support Vector Machine (SVM) in real-world applications is relatively greater than in case of pattern classification, this has been widely investigated in case of machine learning*.**

***Index Terms*: *Sentiment Analysis, local beam search, Genetic Algorithm (GA) and Support Vector Machine (SVM).***

## I. INTRODUCTION

For the purposes of application development as well as hosting, there are a several cloud services that are available. Due to the differences in implementation, every service is unique. When the applications are being run, these differences are apparent, and hence, for determining the suitability of a certain service to the requirements of an application, the evaluations of performance are necessary. The benchmarks for measuring running time, usage of memory, disk read/write operations or other pertinent metrics are comprised in the assessments. The sentiments of an author articulated in either optimistic or pessimistic comments, questions and requests can be analysed across several documents using a Natural language processing and information extraction task referred to as sentiment analysis. In the recent past, the inspiration for sentiment analysis that we see today is the exponential rise in the usage of the internet along with the public opinion. A large amount of both structured and unstructured data are contained in the web and it is challenging to analyse this data for extracting nuanced public sentiments/opinions.

**Revised Manuscript Received on 30 March 2019.**
* Correspondence Author
**P Vasudevan***, Research Scholar, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India.
**K P Kaliyamurthie**, Professor & Dean, Dept. of CSE, Bharath Institute of Higher Education and Research, Chennai, India.



The review of product domain is way different from the data set comprising the movie reviews. In case of the former, as there are some product traits which are preferred and some which are not, the reviewer pens down both positive as well as negative opinion. However, this review cannot be easily classified into positive and negative class. Also, there may be some comments that are specific to features like, the battery lifetime of a laptop being less albeit exhibiting a good holistic performance. It is challenging to identify the overall sentiment of the reviews of this type. In general, there are more comparative sentences in the product review dataset compared to the movie review dataset and this is challenging to categorize [1].

The aspects of the products that are commented by the clients are identified by feature extraction. By classifying the polarity of the sentiment as positive, negative or neutral, the text comprising the sentiment or opinion can be identified by sentiment prediction. Finally, the outcomes that are obtained in the prior step are aggregated using the summarization module. After receiving the text as input, feature extraction process generates the extracted features in different forms such as Lexico-Syntactic or Stylistic, Syntactic and Discourse. The size of the problem decreases, and the classifier performance improves by eliminating the irrelevant features and noise by using the feature selection [2]. It also helps identifying any of the pertinent features for a particular problem. It also aids the learning algorithm performing better. The requirement of computer storage is also lesser with a decrease in the computation time and an increase in the prediction quality due to the reduced features. For tackling high-dimensional data, feature reduction has been proven to be a highly effective data reduction scheme. The subset of pertinent features for the construction of the model is directly chosen by the feature selection. One of the major advantages of feature selection is that it maintains the physical meanings of the original set, providing better interpretation and comprehension of the model, as it retains a subset of original features. The meta-heuristic schemes which mimic the elaborate process of optimization from biological evolutionary process for solving mathematical optimization problems are the Genetic Algorithms (GA). Darwin's 'Survival of the fittest' principle is the motivation for the genetic algorithms.

Problems are considered as abstract individuals in the population. A fitness function is used for evaluating every solution. The possibility of a solution surviving is expressed by the fitness value. This implies that the possibility of being included in the subsequent population and spawning offspring having similar traits by means of genetic trait transmission mechanisms like reproduction, mutation and cross over are all expressed by the fitness values. Mutation of genes and crossover are responsible for reproduction and variation. Variation derives two novel solutions by combining the traits of the two solutions. For the performance of the GA, coding the problem into s genetic denotation, like the sequence of parameters of phenotype on a genotype is very critical. Additionally, there is a huge impact of the fitness function over the performance.

A supervised machine learning algorithm which can categorize the feature space vectors into one of the two sets is the Support vector Machine (SVM) which utilizes the training data from the sets [3]. The operation is performed by building an optimal hyper-plane which can partition the two sets, which is either in the original feature space or a high dimensional kernel space.

This work utilizes the GA based feature selection as well as classification. For this, it makes use of the SVM classifier for sentiment analysis of big data. The remainder of the work has been structured as given below: The second section explains the related work in literature. The third section explains the materials as well as the techniques used. The outcomes are discussed in the fourth section and the fifth section presents the conclusion.

## II. LITERATURE SURVEY

One basic problem of sentiment analysis is categorizing the sentiment polarity and this is covered by Fang & Zhan [4]. A generic scheme for categorizing the sentiment polarity is suggested along with the in-depth description of the process. The online product reviews that are collected from Amazon.com comprise the data used in this study. Experiments for sentence level classification as well as review level classification have been done and the results are effective. Lastly, the insights on the future work in the area of sentiment analysis are described.

The GA for feature selection has been studied extensively by Babatunde et al [5]. In that, specifically, for decreasing the dimensionality, binary GA was used, which could enhance the performance of the associated classifiers .From the image set in the publicly available Flavia dataset, about a hundred features have been extracted. The extracted features are Zernike Moments (ZM), Fourier Descriptors (FD), Lengendre Moments (LM), Hu 7 Moments (Hu7M), Texture Properties (TP) and Geometrical Properties (GP).This article has been mainly contributed by (1) Detailed GA toolbox documentation in MATLAB and (2) Enabling GA to obtain a combinatorial feature set for increased accuracy by developing the GA-based feature selector using a new fitness function (kNN based classification error). The outcomes have been contrasted with several feature selectors from WEKA software. This has helped obtain more accurate outcomes than the WEKA feature selectors from the perspective of classification precision.

For determining the words in a text of document which are favourable to a certain query, Ramos [6] has studies the application of Term Frequency Inverse Document Frequency (TF-IDF). As implied by the term TF-IDF uses an inverse proportion of the number of times with which the word appears in a document to the percentage of the documents comprising the word. It this calculates the values for each word in a document. In case the words have a high TF-IDF numbers, it means that they have a strong association with the documents in which they appear. This means that the document would be of the user's interest if that work were to appear in a query. Evidence has been presented for the efficient categorization of relevant words using this simple algorithm which could enhance the retrieval of query. A novel technique for feature selection as well as sentiment classification has been suggested by Kummer et al [7]. The Z score measure is used for identifying the most salient features that belong to certain categories. Confident features have been identified, based on this score. For obtaining scores for the terms that appear in the neighbourhood of the confident terms, Information Gain (IG) is used. This information is used to propose a novel weighting scheme to perform the sentiment analysis classification. Using different text representation schemes, the suggested feature selection and classification scheme has been analyzed on two datasets that are publicly available. It has been realized that the suggested technique performs the same, sometimes, outperforming schemes like SVM and Naive Bayes, for which accuracy rates across a ten-fold cross validation are used. For finding an optimal feature subset, a combined technique along with two meta-heuristic algorithms is used in Yousefpour et al [8]. There are two steps for performing feature selection: The first is using a hybrid filter and wrapper approaches for obtaining different feature subsets referred to as local solutions, which can decrease the feature space high-dimensionality. The second is integrating local solutions using two meta-heuristic algorithms referred to as genetic algorithm and harmony search algorithm. This helps to find an optimal feature subset. It has been shown by the outcomes on 3 pervasive datasets in sentiment analysis that in terms of accuracy, the suggested scheme performs better than other baseline schemes. A feature selection scheme based on the genetic algorithm (GA) is suggested by Ghareb et al. [9]. By making use of combined search schemes that exploit the benefits of the feature selection as well as the enhanced GA (EGA) in the wrapper scheme, the scheme works. Thus high dimensionality in the feature space can be dealt with. This improves both the crossover and the mutation operators. The basis for the crossover is based on the feature subset which is the chromosome partitioning with term as well as the document frequencies of chromosome entries/features. Based on the significance of the features and the performance of the classifier on original parents, the mutation is performed. Thus, based on the useful information, probability and random selection are supplanted by crossover and mutation operators.

Also, for creating the hybrid feature selection approaches, 6 popular filter selections schemes are incorporated with the EGA. The EGA is applied in the hybrid approach to many feature subsets of varied sizes. These sizes are graded in the descending order based on their importance. For this, reduction in dimension is performed. The EGA operations have been applied to the most significant features having higher ranks. The Naive Bayes approach is used for evaluating the efficacy of the suggested scheme. Also, associative classification is used on three varied collections of the Arabic text datasets. The superiority of the EGA over GA is shown by the empirical outcomes which show that EGA is better in terms of decreased dimensions, time and increased categorization efficacy. Also, the 6 other hybrid FS schemes that comprise filter and EGA techniques are applied on different feature subsets. It has been shown by the outcomes that as these hybrid schemes could produce a better reduction rate without loss of categorization accuracy in most cases, compared to the single filter schemes for reduction in dimensionality. An ensemble approach for feature selection has been suggested by Onan & Korukoğlu [10]. This approach combines many individual feature traits obtained using various filter selection schemes. This is because a stronger and more effective feature subset can be obtained. A genetic algorithm is utilized for combining the individual feature lists. It has been proven by empirical outcomes that the suggested aggregation scheme is effective and performs better than the individual filter-based feature selection schemes for sentiment classification. Keshavarz & Abadeh [11] has suggested a new genetic algorithm for solving this optimization problem. It can also find lexicons for classifying the text. First, adaptive lexicons are generated by the algorithm and then, based on it, a meta-level feature is extracted which is used with Bing Liu's lexicon and n-gram features. The experiments have been performed on 6 datasets. The outcomes have performed better than the popular schemes in literature, in terms of accuracy, on two datasets. Also, the suggested scheme, in four of the datasets performs better in terms of the F-measure. When the suggested scheme is applied on six datasets, the precision increased greater than 80% in all 6 datasets,; in 4 of these sets, the F-measure was higher than 80%. It is possible to deduce the specific language and culture of the Twitter users including the sentiment polarities in varied contexts, by using the sentiment analysis lexicons that are created using the suggested algorithm. It has also been shown that not omitting the traditional stop words is useful. This is because every word will have some sentiment implications.

## III. METHODOLOGY

This work uses a subset of the Amazon book sentiment dataset that comprises 45000 positive, 40000 negative and 35000 neutral datasets. Also, for the extraction of features, a TF based feature extraction has been used. The discussion of GA based selection along with SVM based classifier has been delineated.

### A. Data Set

There are reviews of the product as well as metadata from Amazon. These have one hundred forty two point eight million reviews from May 1996 to July 2014. Links (views and purchased graphs), product metadata (descriptions, category information, price, brand and image features) and reviews (ratings, text, helpfulness votes) are all a part of the dataset.

### B. Term Frequency-Inverse Document Frequency (TF-IDF) Feature Extraction

A numerical measure of the importance of a word to a document in a text or a corpus is referred to as Term Frequency-Inverse Document Frequency (TF-IDF). This formulates a weighing factor for retrieving information and also in text mining. This scheme has been mainly used for halting the filtering of words in text summarization as well as classification schemes. Conventionally, the value of TF-IDF proportionally enhances with the frequency of the word appearing in a document. However, this is nullified by the number of times the word appears in a document. This alludes the occurrence of a few words appearing greater number of times than the others. The raw frequency of a term in the document is referred to as the word frequency. Also, the measure of a term being usual or uncommon across the documents represents the inverse frequency term, which is derived by division of the total quantity of documents to the number of documents comprise the term [12]. TF-IDF, [13] short for term frequency–inverse document frequency, is a numeric measure. This helps scoring the significance of a word in a document. It depends on the number of times a word appears in the document when a collection of documents has been presented. The intuition for this measure is: More often the word appears in a document, more important it is and hence that word deserves a high score. However, in case the word appears in many other documents, it means that it is not unique indenter. Hence the word is assigned a lower score. The math formula for this measure is shown in equation (1):

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

Where the terms are represented by t; d denotes each document; The collection of documents is given by D.

### C. Feature Selection

Selection of a variables subset from the input that can effectively delineate the input data and simultaneously decrease the noise and irrelevant variable affects and still provide effective prediction outcomes is the objective of feature selection.

#### i. Local Beam Search Feature Selection

K of the arbitrarily spawned states form the basis for starting the search procedure. All the successors of the k states are chosen at every step. The algorithm stops in case any one of the successors is the objective. Else, from the complete list, it selects k of the best successors. This process if repeated. When the beam searches are executed in parallel, the failed searches are abandoned and the resources are moved to a place where maximum progress is made.In stochastic beam search the goodness based probability is used for selecting the maintained successor states.

```
OPEN = {initial state}
while OPEN is not empty do
    1. Remove the best node from OPEN, call it n.
    2. If n is the goal state, backtrace path to n (through
        recorded parents) and return path.
    3. Create n's successors.
    4. Evaluate each successor, add it to OPEN, and record
        its parent.
    5. If |OPEN| >ß , take the best ß nodes (according to
        heuristic) and remove the others from the OPEN.
done
```

The disadvantage of local beam search algorithm is:

- All k states can become stuck in a small region of the state space
- To fix this, use stochastic beam search
- Stochastic beam search:

– Doesn't pick best k successors

– Chooses k successors at random, with probability of choosing a given successor being an increasing function of its value.

### ii. Genetic Algorithm (GA) based Feature Selection

The space for the GA to work is the binary search space. This is because the chromosomes are strings of bits. Similar to human evolution, the GA manipulates the finite binary population. Initially, the first population is arbitrarily generated and evaluated using a fitness function A gene value of '1' shows that s certain feature indexed by the position of '1' is chosen , for a binary chromosome employed. If it is '0', the feature will not be chosen for the chromosome to be evaluated. Their grading is done using the positional index of the features indexed by '1's. On the basis of these grades, the top n fittest kids (Elitism of size n) are chosen to be carried over to the subsequent generation.

After automatically pushing the elite offspring to the subsequent generation, the remainder of offspring in the current population go through genetic functional crossover and mutation for respectively forming crossover and mutation offspring. The three offspring which comprise the elite, crossover and mutation, formulate the novel population or new generation. Crossover offspring are generated using a genetic functional crossover process for combining two individuals (chromosomes). Whereas, by means of flipping of bits that depends on its probability, mutation operator is utilized for the genetic disturbance of genes in every chromosome.

For this work, the initial population considered is a matrix of dimensions of population size X length of chromosome. Population size is how many chromosomes are contained in the population. The quantity of bits or genes in each chromosome is known as the length of the chromosome. The size of the population should be at least equal to the value of the length of the chromosome, to enable the individuals in the population to span the whole search space[14].

**Advantages of Genetic Algorithm**

- Parallelism
- Moving in the search space with more individuals and thus there is a lesser probability of getting stuck in local extreme similar to other techniques
- Ease of implementing.

### D. Support Vector Machine (SVM) Classifier

Classification is referred to as the process of categorizing objects into available class. A database at times consists of class or labels of every pattern or instance which is mentioned previously. Supervised classification is the classification based on the class of an unfamiliar pattern found based on familiar patterns. Neural networks and SVM are some of the well-known supervised classifiers. A classifier that can classify the patterns only into two classes is referred to as Support Vector machine. The data classification by the SVM is done by finding the best hyper plane which segregates the data points of one class from the other classes. The best hyper plane for SVM is the hyper plane having the greatest margin between two classes. The maximum slab width parallel to the hyper plane where there are no interior data points is referred to as margin [15]. Data points lie nearest to the segregating hyper plane and these are present on the boundary of the slab. Hyper planes or a set of hyper planes by the SVM are developed in infinite dimension space. The distance from the decision surface to the closest data point determines the classifier margin. Hence, the hyper planes function as decision surface which functions as the condition for deciding the distance of any data point from it. The distance from the closest data point is used for calculating the classifier margin and this leads to a successful classification, although, even a slight error will not lead to misclassification. The benefit of linear SVM is its fast speed of execution and the absence of tuning parameters with the constant as the exception. They are also considerably immune to the comparative sizes of the training samples of the 2 categories. In most learning algorithms, the algorithm will attempt to accurately categorize the class with many examples, if there are several examples of one class than the other. This will help reduce the error rate. The SVMs do not directly attempt to decrease the rate of error, yet strive to segregate the patterns in high dimensional space. The outcome is that the SVMs are comparatively immune to the comparative numbers of every class.

## IV. RESULTS AND DISCUSSION

Table 1 show the parameters used in Genetic Algorithm. Table 2 to 4 and figure 1 to 3 shows the recall, precision and F Measure for Local Beam search feature selection and GA based Feature Selection respectively.

**Table 1 Parameters of Genetic Algorithm (GA)**

| GA Parameter | Value |
|---|---|
| Population size | 100 |
| Genome length | 100 |
| Population type | Bitstrings |
| Fitness Function | kNN-Based Classification Error |
| Number of generations | 300 |
| Crossover | Arithmetic Crossover |
| Crossover Probability | 0.8 |
| Mutation | Uniform Mutation |
| Mutation Probability | 0.1 |
| Selection scheme | Tournament of size 2 |
| Elite Count | 2 |

**Table 2 Recall for GA based Feature Selection**

| Training percentage | Local Beam Search Feature Selection | GA based Feature Selection |
|---|---|---|
| 20% | 22.2694 | 21.32617 |
| 40% | 23.68047 | 23.39307 |
| 60% | 23.95713 | 25.66833 |
| 80% | 24.92833 | 26.8839 |

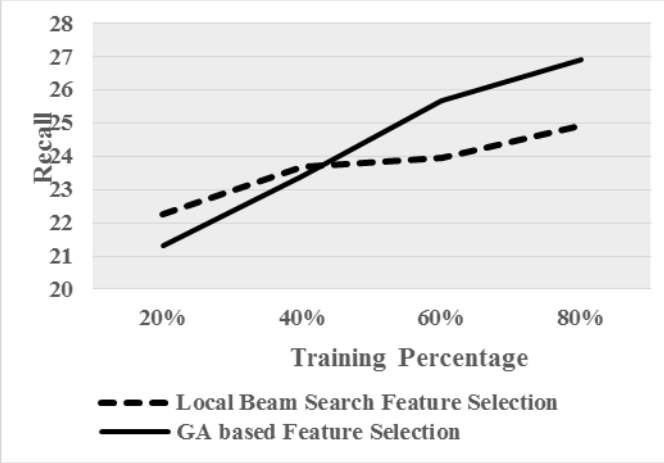


**Figure 1 Recall for GA based Feature Selection**

Table 2 and Figure 1 shows that the recall of GA based Feature Selection performs lower by 4.33%, by 1.22%, better by 6.89% and by 7.55% at the training percentage 20%, 40%, 60% and 80% respectively than Local Beam Search Feature Selection.

**Table 3 Precision for GA based Feature Selection**

| Training percentage | Local Beam Search Feature Selection | GA based Feature Selection |
|---|---|---|
| 20% | 0.6633 | 0.6327 |
| 40% | 0.691533 | 0.689433 |
| 60% | 0.706433 | 0.7433 |
| 80% | 0.724233 | 0.782033 |

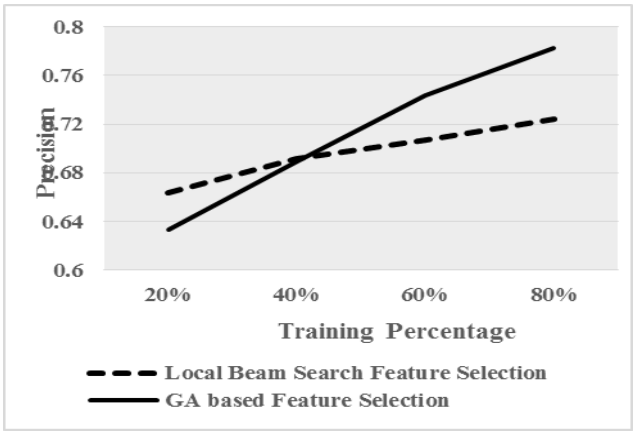


**Figure 2 Precision for GA based Feature Selection**

**Table 4 F Measure for GA based Feature Selection**

| Training percentage | Local Beam Search Feature Selection | GA based Feature Selection |
|---|---|---|
| 20% | 0.647767 | 0.620567 |
| 40% | 0.695567 | 0.684433 |
| 60% | 0.701267 | 0.7557 |
| 80% | 0.731733 | 0.789767 |

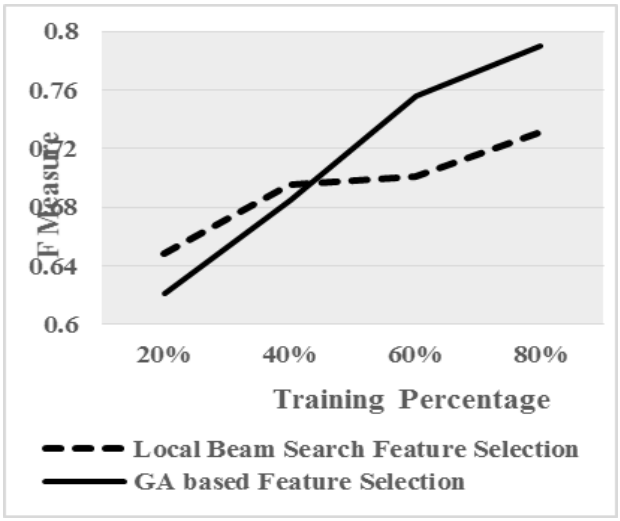


**Figure 3 F Measure for GA based Feature Selection**

Table 3 and Figure 2 shows that the precision of GA based Feature Selection performs lower by 4.72%, by 0.3%, better by 5.09% and by 7.67% at the training percentage 20%, 40%, 60% and 80% respectively than Local Beam Search Feature Selection.

Table 4 and Figure 3 shows that the F Measure of GA based Feature Selection performs lower by 4.29%, by 1.61%, better by 7.5% and by 7.63% at the training percentage 20%, 40%, 60% and 80% respectively than Local Beam Search Feature Selection.

## V. CONCLUSION

An overview of the experience of an individual or his opinion about a product is provided by sentiment analysis. This task is challenging owing to the huge amount of online data.

Irrelevant features are eliminated and important features are selected by the feature selection schemes. The computation speed is improvised by reduced feature vector which comprises pertinent features. This also improves the precision of the machine learning techniques. This work proposes genetic algorithm which is based on feature selection. These genetic algorithms are based search method which can be used to optimize a set of parameters in a search space. Searching a decision boundary between two classes which is located at a distance from any point in the training data is the basic objective of the support vector machine. Outcomes have demonstrated that the precision of GA based Feature Selection performs lower by 4.72%, by 0.3%, better by 5.09% and by 7.67% at the training percentage 20%, 40%, 60% and 80% respectively than Local Beam Search Feature Selection.

## REFERENCES


1. Agarwal, B., & Mittal, N. (2013, March). Optimal feature selection for sentiment analysis. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 13-24). Springer, Berlin, Heidelberg.
2. Kumar, K., & Kumar, G. Analysis of Feature Selection Techniques: A Data Mining Approach.
3. Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. Physical review letters, 113(13), 130503.
4. Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 5.
5. Babatunde, O. H., Armstrong, L., Leng, J., &Diepeveen, D. (2014). A genetic algorithm-based feature selection.
6. Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).
7. Kummer, O., Savoy, J., & Argand, R. E. (2012). Feature selection in sentiment analysis.
8. Yousefpour, A., Ibrahim, R., Hamed, H. N. A., & Yokoi, T. (2016, March). Integrated Feature Selection Methods Using Metaheuristic Algorithms for Sentiment Analysis. In Asian Conference on Intelligent Information and Database Systems(pp. 129-140). Springer, Berlin, Heidelberg.
9. Ghareb, A. S., Bakar, A. A., &Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications, 49, 31-47.
10. Onan, A., &Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. Journal of Information Science, 43(1), 25-38.
11. Keshavarz, H., &Abadeh, M. S. (2017). ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. Knowledge-Based Systems, 122, 1-16.
12. Munot, N., &Govilkar, S. S. (2014). Comparative study of text summarization methods. International Journal of Computer Applications, 102(12), 33-37.
13. Khan, A. U., Bandopadhyaya, T. K., and Sharma, S.: Comparisons of Stock Rates Prediction Accuracy using Different Technical Indicators with Backpropagation Neural Network and Genetic Algorithm Based Backpropagation Neural Network. In: Proceedings of the First International Conference on Emerging Trends in Engineering and Technology IEEE Computer Society, Nagpur, India. (2008).
14. Mathworks T. Statistics Toolbox User's Guide The MathWorks, Inc. 3 Apple Hill Drive Natick, MA 01760-2098, 2013.
15. Dubey, V. K., & Saxena, A. K. (2016, March). Hybrid classification model of correlation-based feature selection and support vector machine. In Current Trends in Advanced Computing (ICCTAC), IEEE International Conference on (pp. 1-6). IEEE.