JACET Journal of Advances in Computer Engineering and Technology, 1(2) 2015

Improving the Operation of Text Categorization Systems with Selecting Proper Features Based on PSO-LA

Mozhgan Rahimirad[1], Mohammad Mosleh[2] and Amir Masoud Rahmani[3]

Received (2015-01-19)
Revised (2015-01-22)
Accepted (2015-03-12)

Abstract — **With the explosive growth in amount of information, it is highly required to utilize tools and methods in order to search, filter and manage resources. One of the major problems in text classification relates to the high dimensional feature spaces. Therefore, the main goal of text classification is to reduce the dimensionality of features space. There are many feature selection methods. However, only a few methods are utilized for huge text classification problems. In this paper, we propose a new wrapper method based on Particle Swarm Optimization (PSO) algorithm and Support Vector Machine (SVM). We combine it with Learning Automata in order to make it more efficient. To evaluate the efficiency of the proposed method, we compare it with a method which selects features based on Genetic Algorithm over the Reuters-21578 dataset. The simulation results show that our proposed algorithm works more efficiently.**

Index Terms — **Text mining; feature selection; classification; Learning Automata(LA); Particle Swarm optimization(PSO).**

I. INTRODUCTION

The text databases are growing quickly due to the increase of digital information. Text Mining as a branch of data mining utilizes data mining techniques over text data. Discovering useful knowledge from semi-structured or unstructured text is one the problems in text mining which is attracted more attentions recently. The goal of text mining is to automate most of the work which is done by users traditionally. One of the major parts of text mining is text classification. In the text classification, it is common to use the words of text as the features. So, the text classification methods conquer a lot of features. In order to reduce the number of features and select related ones, many different methods are introduced in the literature. This is done before the classification phase and it aims to make the dataset simpler by reducing the dimensionality and unifying the basic related features without sacrificing the original data by predictions [1]. Feature selection is extended into many fields such as: text classification, data mining, pattern recognition and signal processing [2].

In the current research, a new method is proposed using Particle Swarm Optimization (PSO) algorithm based on the learning automata to solve the problem of feature selection. The proposed method for feature selection is a wrapper method and the binary PSO evolutionary algorithm is utilized to search the problem space. In order to improve the efficiency of our algorithm, it is combined with object migrating learning automaton which has a fixed structure. This brings a better feature selection using the reward and penalty system used in automata. Subsequently, the extracted features are given to

[1] Department of Computer, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran (Mozhgan_rahimirad@yahoo.com).
[2] Department of Computer Engineering, Dezfoul Branch, Islamic Azad University, Dezfoul, Iran (mosleh@iaud.ac.ir).
[3] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran (rahmani@srbiau.ac.ir).

the SVM and KNN classifiers and the accuracy of the classifier is measured for each classifier separately.

The remaining of this paper is organized as follows. The related works are provided in section 2. In section 3, the proposed feature selection algorithm and its steps are described in details. Section 4 depicts the simulation results of comparison among the proposed algorithm and Genetic algorithm. In section 5, we conclude our paper and finally we provide some offer for future works to improve the proposed algorithm in section 6.

II. RELATED WORKS

Most samples have many features which can be demonstrated with vectors in a high dimensional space. The goal of feature selection is to reduce the dimensionality of data as it is possible. There are many reasons for dimensionality reduction. It removes excess and irrelevant features and also it improves the efficiency of classifier [2]. The further analysis of the classification results will be simpler and also it helps to understand the problem easier. A lot of methods are proposed in the literature to improve the feature selection procedure. In [3], a two steps feature selection method for text classification is introduced. In the first step, a new feature selection method is applied to reduce the less important terms and in the second step, a new semantic space creates a hidden semantic between words according to the indexing method. In [4], a new method for document classification is introduced which uses feature extraction based on the graph. In this method, a set of documents (as the set of graphs) are sent to the weighted graph extraction algorithm and subsequently, the major sub-graph is extracted. In [5], some metrics are introduced to select a feature subset for Arabic context classification. In this work, an experimental comparison is done between 17 metrics for traditional feature subset selection methods and the results show that the Chi-square and Fallout feature subset selection work more efficiently for Arabic text classification purposes. A filter model for feature subset selection based on Genetic algorithm is proposed in [6]. This model describes a new feature subset selection algorithm which uses the Genetic Algorithm (GA) to optimize the output nodes of the learned Artificial Neural Network (ANN). In [7] the divergence based on feature selection is proposed for different classes. Most of the text feature selection methods only aim at auditing the relation and redundancy analysis from the class viewpoint. However, it ignores the relation between different classes and features. In [8], a two steps feature selection and feature extraction is utilized in order to improve the efficiency of text classification. During the text classification, words of less importance are ignored and the feature selection and feature extraction procedures are used only for words of higher importance. In that way, the computation time and classification complexity is reduced. A novel feature selection algorithm based on Ant Colony Optimization is proposed in [9] to improve the efficacy of text classification. The Ant Colony Optimization algorithm is inspired from the behavior of ants. They can find the minimum length path to the food sources after some searches. The algorithm executes easily and due to the use of simple classifier, the computation complexity is reduced in a significant way. In [10], feature selection strategies used for text classification are compared together and evaluated using popular feature selection metrics. Moreover, a framework namely ‘comparative keyword selection’ is proposed in order to select the keyword.

Among many methods proposed for feature selection, the Genetic and Ant Colony algorithms work more efficiently. These algorithms find the best solutions according to the knowledge obtained from previous iterations.

III. THE PROPOSED METHOD

All of the previous methods tried to reduce the feature dimensionality in order to improve the accuracy and speed of classifiers and each of them were succeeded approximately. Evolutionary algorithms are more suitable for large datasets due to the fact they can search the problem space and find the best solution more quickly. The basic idea of this research is to design and implement a system with minimum number of features and acceptable efficiency to improve the accuracy, efficiency and computation cost of the classifier.

In the current research, a new method is proposed using Particle Swarm Optimization (PSO) algorithm based on the learning automata to solve the problem of feature selection. The proposed method for feature selection is a wrapper method and the binary PSO evolutionary algorithm is utilized to search the problem space. In order to improve the efficiency of our algorithm, it is combined with object migrating learning automaton which has a fixed structure. This brings a better feature selection using the reward and penalty system used in automata. Subsequently, the extracted features are given to the SVM and KNN classifiers and the accuracy of the classifier is measured for each of them using classification metrics separately. Fig.1 depicts the diagram of proposed method:

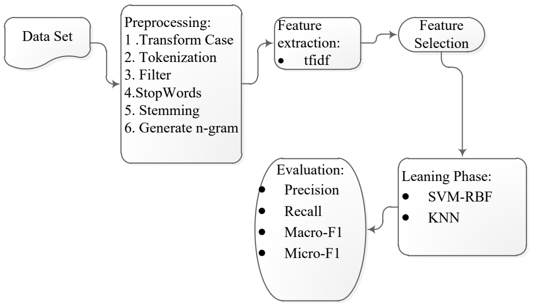

Fig. 1.Block diagram of the proposed improvements to the text mining

1. Text Preprocessing

In order to classify texts, a preprocessing phase is applied after the dataset is collected. In this step, the raw text documents should be converted in to appropriate form which can be used in feature selection and learning steps. The dataset preprocessing consists of the following steps:

- Transform case: in this step, all of the characters are converted in to the lower-case letters.
- Tokenization: in this step, the entire text is divided in to separate consecutive words.
- Filter StopWords: in this step, the extra words of less importance in English language are removed.
- Stemming: in this step, the words stems are determined using the Porter Stemmer algorithm to strip the prefix and suffix of the words and decreases the length of the words until the stem forms are obtained.
- Generate n-gram: in this step, we utilize n-gram to index and reduce the text dimensionality. According to the n-gram, we can demonstrate a typical text as a series of consecutive words with the length of n. At the first, this model is introduced for speech processing problems. However, there are different versions of this model generalized for natural language problems and text classification [11]. In order to prevent complexity and according to experiments performed over different values of n, we use n-gram considering n=2.

2. Feature Extraction

The feature extraction based methods, map a multidimensional space to a lower dimensionality space. In the current research, we utilize a weighting method namely tfidf. The efficiency of the classifiers is evaluated using this weighting method [12].

3. The proposed feature selection algorithm

In this step, the proposed method uses the Particle Swarm Optimization and Learning Automata in order to select features. We describe the details of this method as follows:

1. Start
2. Production and initialize the particles with random values for velocity and position
3. Calculate the fitness function for the particles (the particles Position)
4. If the particles are evaluated from the local memory is better to replace it now
5. If the particles are evaluated to be better than the best memory, replace it now
6. speed update now.
7. position update now.
8. Apply mutation on particle
9. Some particles are given to automata randomly. Subsequently, the fitness function is evaluated for these particles. We refer to this value as f1.
10. For some particles, the bits are changed randomly (from 0 to 1 or vice versa) and the fitness function is evaluated again. The new value is referred to as f2.
11. If f1>f2 the automaton is rewarded. Otherwise, it is penalized.
12. If the stop condition is satisfied the algorithm terminates. Otherwise, the process is repeated from step 2.
13. End

The steps of feature extraction procedure are as follows:

We utilize the binary PSO algorithm. Each particle is demonstrated by an automaton. The velocity and position parameters are defined for each particle. Here the positions are the set of actions in automata. Moreover, the local best position of the particle is stored too. Also, the global best position among all particles is determined and stored. These values are used to determine the new position of each particle in the next iterations. The importance coefficients of these values (local and global best solutions) are determined using c1 and c2. The sum of these coefficients must be equal to 4 (this value is obtained experimentally and is considered in all PSO implementations). Each feature is assigned to one of the automata actions and stand in a specific depth of automata (the total number of actions equals to the total number of features). In other words, each action of automata is considered equivalent to a feature. We assign a 'one' value to an action (there exist a feature in the set of solution) or 'zero' value (there is no feature in the solution set). Each automaton has 5 states for each action. The value of corresponding action can be resolved according to these states and each feature is classified based on its state. Here we consider two states: internal and boundary. According to reward or penalty of an action, the state of corresponding feature is changed. On reward, the corresponding feature moves to its internal state and on penalty it moves towards the boundary state. Consider a feature which is on the boundary state. If it is penalized, the selection status of corresponding feature is changed and therefore a new composition for solution set is generated. The set {1, 0, 0, 1, 0, 1} depicts a set of features. It demonstrates that the features labeled with 1, 4 and 6 are chosen while the ones labeled with 2, 3 and 5 are not.

One of the solutions which help to conquer the local optimum problem in particle swarm movements is to utilize mutation [13]. Here we use the Cauchy distribution for the mutation and each particle is mutated according to the probability Pmutate. Once a particle is selected for mutation, each of the elements of its vector is mutated with probability 1/d where d is the problem dimensionality. In order to mutate each element of a particle, a random number is generated using the Cauchy distribution. This number is summed up with the value of corresponding element of particle. The proposed method gives an appropriate solution for some of the problems with large number states.

- Fitness Function
  In order to evaluate the eligibility of each particle, the selected features are given to the KNN classifier and data are classified. Here, the fitness value for each particle is considered equal to the classifier accuracy.
- Reward and Penalty procedures in Automata
  After fitness evaluation for each action of Automata, the corresponding feature is rewarded or penalized: first the fitness value is obtained for the particle. This value is referred to as f1. After the next movement, the fitness value is recomputed for the same particle and referred to as f2. If the value of f1 be greater than f2, the action is rewarded, otherwise it is penalized. In the case of reward or penalty for an action, the state of the feature changes in the set of states for corresponding action. If a feature in the boundary state is penalized, the value of the corresponding action is changed and therefore, a new set of solution is generated. The rate of this operator should be low because it is a random search operator and applying it with higher rates decreases the efficiency of the algorithm.
- The penalization of an action can be done according to the current state of the relevant value:

1. The relevant value is in a non-boundary state: the penalization moves the relevant value towards the boundary state. In other words, if the relevant value equals to 1, the penalization means that choosing this feature probably is not an appropriate choice. If the relevant value equals to 0, penalization means that removing this feature from the solution was not an appropriate task.
2. The relevant value is in the boundary state: in this case, penalization changes the relevant value.

In order to update the local best solution, by using (1):

If fitness (x)> fitness (pbest)
Pbest=x (1)

In order to update the global best solution, by using (2):

If fitness (x)> fitness (Gbest)
Gbest=x (2)

In each iteration of the search, all members are updated considering two best values. The first one relates to the best solution ever found by the bird up to this time (the eligibility value of this solution is stored too). This value is referred to as Pbest. The second best which is tracked by the PSO is the best position found by the population heretofore. This is a global best value and is referred to as Gbest. In the case that a member considers part of the population as its neighboring topology, the best value is local best and is referred to as Lbest. After these values are determined, the velocity and position of each member are updated using (3) and (4):

$$v_i^d = w * v_i^d + c_1 r_1 * \left(pbest_i^d - x_i^d\right) + c_2 r_2 * \left(gbest^d - x_i^d\right) \qquad (3)$$

$$\left. sig\left(v_i^d\right) = \frac{1}{1 + e^{-v_i^d}} \right\} \; if \; sig\left(v_i^d\right) > r_3 \rightarrow x_i^d = 1 \;\; else \;\; x_i^d = 0 \qquad (4)$$

In these equations, i represents the number of iteration and variables c1, c2 represent the learning factors. Often we consider c1 = c2 = 2 in order to control the movement size of a bird at once. r1 and r2 are two uniform random numbers in [0, 1]. w is an algebraic weight which is typically initialized in [0, 1]. A larger algebraic weight simplifies the global exploration and a smaller algebraic weight simplifies local exploration. In the standard PSO algorithm, the population is initialized by random solutions. The population fitness values are evaluated and the Pbest, Gbest, Velocity and positions are updated until the stop condition is satisfied. Finally, the Gbest and its fitness value are demonstrated as the output.

4. The Leaning Phase

The features chosen by the proposed method are given to the KNN and SVM-RBF classifiers to evaluate the efficiency of our algorithm. This indicates how efficiently our method improves the accuracy and speed of classifiers in comparison with state of the art methods.

IV. SIMULATION RESULTS

Some experiments are performed to demonstrate the efficacy of the proposed algorithm. All experiments are performed over a system running Microsoft Windows, an Intel Core i7 1.60 GHz CPU and 4 GB main memory. We implemented the proposed method using Matlab R2013a. Also, for comparison purposes, the Genetic Algorithm is implemented for feature selection [14] and is compared with our algorithm.

1.Dataset

The dataset used in this research is the Reuters-21578 news dataset. Details of the used dataset are depicted in the following table I:

TABLE I
Characteristics Of Data Sets Reuters -21,578, Subset R (8)

categories	Number of training documents	The total number of documents in categories	Number of Test Documentss
Acq	1596	2292	696
Crude	253	374	121
Earn	2840	3923	1083
Grain	41	51	10
Interest	190	271	81
Money-fx	206	293	87
Ship	108	144	36
Trade	251	326	75

2. Parameters of the proposed method

TABLE II
Parameters Of The Proposed Method

Amount	Parameter	Signs
1.5	Best global Factor	C1
2.5	Best global Factor	C2
[0,1]	Weight algebraic	W
[0,1]	Random number	R1
[0,1]	Random number	R2
0.3	Mutation probability	Mutation_p
0.2	Probability of reward and punishment	Rp_p
100	The maximum number of iterations of the algorithm	Max_it
5	Status	La_depth
1.5	Best global Factor	C1
2.5	Best global Factor	C2
[0,1]	Weight algebraic	W
[0,1]	Random number	R1
[0,1]	Random number	R2
0.3	Mutation probability	Mutation_p

3. Evaluation of proposals

In this step, the models are implemented and evaluated using the mentioned metrics which are: Precision average, recall average and F1 average. Note that all of the obtained results are depicted in percentage.

- Precision: The Precision metric depicts the percentage of documents classified correctly into a class to the total number of documents classified to the same class. In other words, it demonstrates the classification accuracy of class i in accordance to the all cases which are labeled with i by the classifier. This metric can be computed , by using (5). Note that the index i in these parameters means that all of the parameters should be evaluated for each class i.

- Recall: The recall metric for a class, demonstrates the percentage of text documents classified correctly among all of the documents which belong to that class. In other words, it demonstrates the classification accuracy considering the total samples labeled with i. The recall metric can be evaluated , by using (6).

$$\mathrm{Precision_i} = \frac{\mathrm{TP_i}}{\mathrm{TP_i} + \mathrm{FP_i}} \qquad (5)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \qquad (6)$$

where TPi is the number of test documents correctly classified under ith category (ci), FPi is the number of test documents incorrectly classified under ci. [9]

The major point is that the recall metric demonstrates the efficacy of classifier according to occurrences of i while the Precision metric is based on the classifier prediction accuracy and demonstrates how we can rely on the output of classifier.

The F1 metric is obtained by combining previous metrics and can be used in the cases where there is no importance or priority among Precision and Recall metrics. This metric can be computed, by usin (7).

$$F1_i = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \qquad (7)$$

SVM_RBF classifier is used here to learn [15]. Table III shows the results of SVM classifier.

Fig.2 and Fig.3 show the average of micro-f1 and macro-f1 metrics for each feature selection algorithms using the SVM classifier. This indicates that the proposed algorithm is optimized more quickly.

TABLE III
Comparison Of The Model With GA Model Using SVM Classifier

Category Name	GA			PSO+LA		
	Precision	Recall	F1	Precision	Recall	F1
Acq	49.72	89.48	63.92	71.59	95.78	81.94
Crude	19.22	72.35	30.37	14.42	18.34	16.15
Earn	95.64	67.11	78.87	99.18	92.35	95.64
Grain	100	100	100	100	100	100
Interest	100	18.1	30.65	100	13.39	23.62
Money-fx	100	18.52	31.25	100	15.36	26.63
Ship	76.08	100	86.42	78.18	100	87.75
Trade	80.72	58.26	67.68	68.82	59.54	63.84
Average	77.67	65.48	61.15	79.02	61.85	61.95

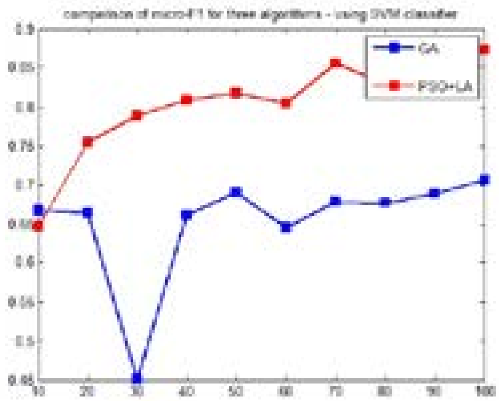

Fig.2. Comparison of Micro-F1 of the method using SVM classifier

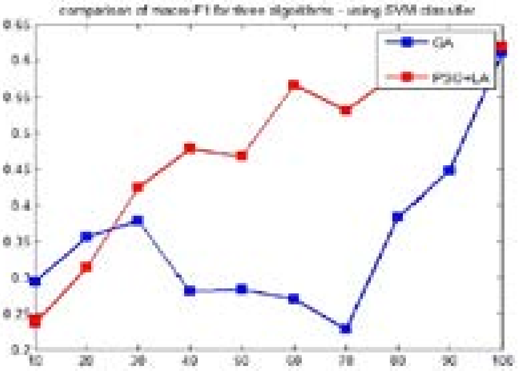

Fig.3. compares the Macro-F1 of the method using SVM classifier

At this stage the KNN classifier is used for learning [11]. Table IV shows the results of KNN classification.

Fig.4 and Fig.5 show the average of micro-f1 and macro-f1 metrics for each feature selection algorithms using the KNN classifier. This indicates that the proposed algorithm works more efficiently in comparison with GA.

Table IV
Comparison of the model with GA models using KNN classification

Category Name	GA			PSO+LA		
	Precision	Recall	F1	Precision	Recall	F1
Acq	59.09	93.72	72.48	79.77	100	88.75
Crude	43.2	43.2	43.2	64.36	47.45	54.63
Earn	93.85	81.66	87.33	100	95.39	97.64
Grain	100	100	100	100	100	100
Interest	100	19.96	33.27	100	26.53	41.93
Money-fx	100	8.81	16.19	100	100	100
Ship	78.32	100	87.84	79.47	100	88.56
Trade	68.78	30.59	42.34	76.7	45.15	56.84
Average	80.41	59.74	60.33	87.54	76.82	78.54

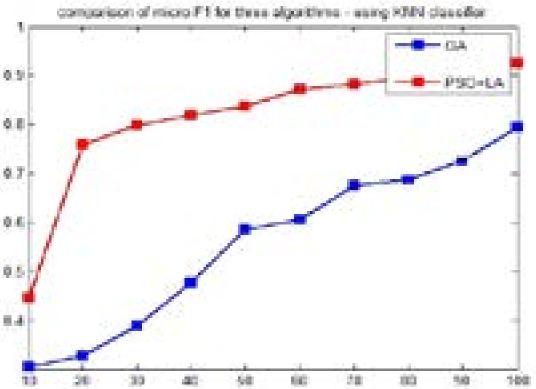

Fig.4. Comparison of Micro-F1 of classification algorithm using KNN.

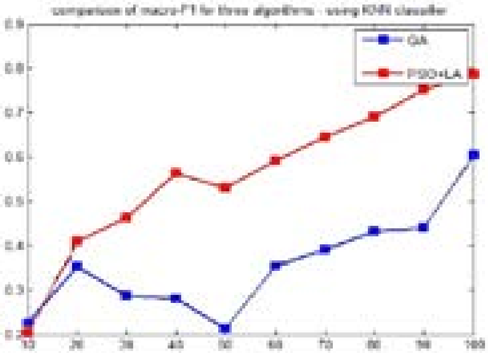

Fig.5. Comparison of Macro-F1 of classification algorithm using KNN.

V.CONCLUSION

The goal of this research is to propose a model for feature selection using PSO evolutionary algorithm and the learning automata in order to improve the accuracy of classifier. Text classification is consisted of two main parts: Feature Selection and Learning algorithm. For the feature selection, wrapper methods are utilized. We used binary PSO evolutionary algorithm to search the problem space and in order to improve the efficiency, it is combined with learning automata. This helps to select better features using the reward and penalty system of automata. Subsequently, the extracted features are given to the SVM-RBF and KNN classifiers separately and the accuracy of classifier is measured using: precision average, recall average and F1 average. The evaluation results indicate that the proposed method increased the accuracy of classifier in comparison with others.

The proposed method surpass previously introduced methods from three points of view: 1- It works more efficiently over high dimensional datasets and its efficacy is not affected with the increase of features. 2- The PSO algorithm has fewer operators in comparison with other evolutionary algorithms and therefore the implementation is so simpler. 3- There is an information stream among the particles.

VI. RECOMMENDATIONS AND FUTURE WORK

The model presented in this paper for feature selection was a wrapper method. As a suggestion for future work will be performed the following tasks:

- In order to improve the classifier we can take advantage of the ability to LA for a different classifier.
- A feature selection method based on the combination of the filter and LA wrapper provided.
- The proposed method can be used in the classification.
- The proposed method can be used for data collection, such as datasets Persian News from Hamshahri News collection is implemented.

REFERENCES

[1] Guyon, I., Elisseeff, A, “An introduction to variable and feature selection”, Journal of Machine Learning Research, Vol. 3, pp. 1157-1182. 2003.

[2] Jensen, R, “Combining rough and fuzzy sets for feature selection,” PhD Thesise, University of Edinburgh, UK. 2005.

[3] Jiana Meng, Hongfei Lin , Yuhai Yu , “A two-stage feature selection method for text categorization,” Computers and Mathematics with Applications 62 pp: 2793–2800. 2011.

[4] Chuntao Jiang, Frans Coenen, Robert Sanderson,

Michele Zito, “Text classification using graph mining-based feature extraction,” Knowledge-Based Systems 23 pp: 302–308. 2010.

[5] Abdelwadood Moh’d Mesleh, “Feature sub-set selection metrics for Arabic text classification,” Pattern Recognition Letters 32 pp: 1922–1929. 2011.

[6] M.E. ElAlami, “A filter model for feature subset selection based on genetic algorithm,” Knowledge-Based Systems 22 pp: 356–362. 2009.

[7] Yishi Zhang, ShujuanLi, TengWang, ZigangZhang, “Divergence-based feature selection for separate classes,” Neurocomputing 101 pp:32–42. 2013.

[8] Harun Ugˇuz, “A two-stage feature selection method for text categorization by using information gain,” principal component analysis and genetic algorithm , Knowledge-Based Systems 24 pp: 1024–1032. 2011.

[9] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri, “Text feature selection using ant colony optimization,” Expert Systems with Applications 36 pp:6843–6853. 2009.

[10] S _erafettin Tas _cı, TungaGüngör, “Comparison of text feature selection policies and using an adaptive framework,” Expert Systems with Applications 40 pp: 4871–4886. 2013.

[11] Sebastiani, F,“Machine Learning in Automated Text Categorization”, ACM Computing Surveys, Vol. 34, No. 1, pp. 107-131. 2002.

[12] Wei, Z., Miao, D., Hugues, J., Zhao, R., Li, W., “N-grams based feature selection and text representation for Chinese Text Classification,” International Journal of Computational Intelligence Systems, 2 (4), pp. 365-374. 2009.

[13] Lan, M., Tan, C. L., “Supervised and Traditional Term Weighting Methods for Automatic Text Categorization,” Journal of IEEE Pami, 10 (10), pp. 1-36. 2007.

[14] M.E. ElAlami,. “A filter model for feature subset selection based on genetic algorithm,” Knowledge-Based Systems 22 pp: 356–362. 2009.

[15] Khan, A., Baharudin, B., Lee, L. H., Khan, K.,. “ A Review of Machine Learning Algorithms for Text-Documents Classification,” Journal of Advances in Information Technology, 1(1), pp. 4-20. 2010.