International Journal of Engineering & Technology, 7 (2.27) (2018) 156-160

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An efficient technique for hybrid classification and feature extraction using normalization

Bipanjyot Kaur [1] *, Gourav Bathla [2]

[1]M.Tech Scholar, department of Computer Science Engineering, Chandigarh University, Gharuan, Mohali, Punjab India
[2] Assistant professor, department of Computer Science Engineering,Chandigarh University, Gharuan, Mohali, Punjab India
**Corresponding author E-mail:bipanjyotkaur@gmail.com*

Abstract

Text classification is technique for assigning the class or label to a particular document within predefined class labels. Predefined classes examples are sports, business, technical, education and science etc. Classification is supervised learning technique i.e. these classes are trained with certain features and then document is classified based on similarity measure with these trained document set. Text classification is used in many applications like assigning the label to the documents, separating the spam messages from the genuine one, filtering of text, natural language processing etc. Feature selection, extraction and classification are various phases for assigning label to any document. In this paper, PCA is used for feature extraction, ABC is used for feature selection and SVM is used for classification. PCA is improved by applying normalization-using size of features in our proposed approach. It reduces the redundant features to larger extent. There are very few research works, which have implemented PCA, ABC and SVM for complete classification. Evaluation parameters like accuracy, F-measure and G-mean are calculated to check classifier efficiency. The proposed system is deployed on 20-Newsgroup dataset. Experiment analysis proves that accuracy is improved using our proposed approach as compared to existing approaches.

Keywords: *TextMining; Text Classification; Feature Extraction; Feature Selection; Machine Learning*

1. Introduction

There is vast amount of data generated in business organizations and social networks. Data is of heterogeneous variety and belongs to different classes. If any user search any topic, it is very difficult to analyze all classes of data. Classification is used to assign these documents a particular class. User search can be checked only for the required class. Classes are chosen from a formerly settled scientific classification (an order of categories or classes). The objective of text classification is to allot reports, (e.g. posts, messages, item surveys, emails, etc.) to several classes. The main issue in text classification is large dimension of data. Feature extraction and feature selection is widely used to reduce the complexity of the dimensionality of data. Text Mining is a process of analyzing and extracting the hidden / unknown but useful data out of large amount of data.[1] An automatic approach to assign label or class to the documents that are given with predefined set of topics or classes, which is refer to as automatic Text classification (TC)[2][3].Various text classifiers are used like decision tree, SVM, Naïve Bayes, ANN etc[2][3][4]. That is used by researchers in their research work. There are various performance parameters such as accuracy, F-measure ,G-mean, precision and recall. There are various types of classification techniques such as single label and multi label classification. In single label only one class is assigned to the document where as in multi label more than one class can be assigned to the document. In this paper, single classification is implemented.

Text classification is used in various fields [2], [5].

i) To define a class of document.
ii) Spam filtering (separation of the fake email from genuine)
iii) Automatic indexing
iv) To determine language of text.
v) Semantic analysis.
vi) Text filtering and so on.
vii) Used for processing of big data.

In this paper, classification is done using machine learning [6] technique. It works in two phases- training section and testing section. In the training section the dataset is trained whereas in the testing section the efficiency of the classifier is evaluated that with how much accuracy it identifies the category of a document.

Our proposed approach includes PCA that tries to reduce the complexity of dimension of features. In our approach, PCA is improved using normalization [7], [8] with by dividing it total no of observed features and ABC and SVM will improve the efficiency of the classification Selection of the features is done through the artificial bee colony [9]. The optimized features are deployed on support vector machine (SVM) [10], which results in better classification.

2. Literature survey

Zobeidi et al. [11] stated a novel approach of classification in which PCA is used for the feature extraction and finally for classification ,MLF is used . The data sets used were 20 newsgroup and Reuters-21578. in this paper accuracy is 95% and f-`measure is 0.88.The limitation of approach is accuracy and F-measure is not upto mark. Tang et al. [12] stated the Naïve Bayesian classification technique for categorization of text using class-specific characteristics. In this a feature subset is chosen from every class. Baggentoss pdf theorem was used to redevelop the PDFs in raw data space from class specific PDFs to build the Bayes classification rule the importance of given approach is that feature selection

criteria, like: MD (Maximum Discrimination), IG (Information Gain) are included easily. Evaluated the performance on several actual benchmark dataset and compared with feature selection approaches. Authors have tested approach for texture classification on binary real time benchmarks: 20- Reuters and 20-Newgroups. The limitation of the approach is different classifiers like ANN, SVM etc. can be used to enhance classification performance [13] describes that as the documents in the digital forms are increasing so the text classification is requires to separate them. The efficiency of the text classification can be enhanced by the feature extraction and feature selection that is done before classification. This paper describes the comparative analysis of various feature extraction and feature selection techniques. This paper describes different classifiers to classify texture documents. Vijayan et al. [14] describes the text classification and the various classification algorithms. Larger amount of the text is stored in the form of the web documents. Classification is describes as the process of classifying text documents in specific number of the predefined classes It also states the various applications of the such as spam filtering ,sentimental analysis ,identification of language etc. Uzer et al. [15] describes a composite approach for classification using ABC and SVM algorithm. The proposed approach was tested on a UCI database i.e diagnosis of the different diseases liver disorders, diabetes dataset. The proposed approach was able to achieve the accuracies of 94.92, 74.81, 79.92.the limitation of proposed approach is accuracies are not up to mark. Santoso et al. [16] states that Internet consist of the larger amount of the unstructured as well as unorganized data. So, the proposed approach used the Naïve Bayes classifier and map-reduce algorithm. This approach was tested on the larger datasets. This approach improved the accuracy of the classification. This approach is robust and classify the document with the good accuracy in their specific domain. The limitation of the approach is as only performance metrics used are few. Xue et al. [17] proposed an algorithm SABCGB, which is self-adaptive depending on best candidate for global candidate. Smart algorithms are depending upon innovative swarm principles actively researched recently. ABC algorithm is an active and smart calculation for issues in global optimization. ABC algorithm is robust and effective optimization technique. Although the search equation utilized in ABC is deficient and candidate solutions generating method's exploring ability is good but exploitation performance is poor. Few complex methods are recently developed for candidate solution, but the robustness and universality are still deficient. For fair contrast with other algorithms, they have used identical initial population in the calculation on benchmark function. The proposed approach was tested on 25 benchmarks. Somvanshi et al. [18] presents the survey of the classification techniques that are also used in the machine learning that is the decision tree and the support vector machine It states that the decision tree is useful for the classification of discretionary data, whereas the SVM is useful to classify both linear and non linear data. SVM is also given importance when multi label data has to classified dataset. In Demidova et al. [19], authors present a novel approach for the selection of the optimal parameters for SMOTE algorithm. In this the SVM classifier is used for classification of the imbalanced datasets. The proposed approach helps to reduce the time needed for the selection of the optimum parameters and enhance the classification performance. Wu et al. [20] states that the social media is used for the gaining the abundant information to study human behaviour, opinions as well as emotions for the events such as natural disasters. This paper uses hurricane sandy 2012. It uses the twitter messages for the information retrieval as and classification of the text. The approach used for the classification of the text is the fuzzy based logic. In this input were the multiple features which were extracted from the twitter's message, whereas the output shows the relevance or importance of the message. Several fuzzy rules was designed and de-fuzzification techniques were combined to obtain the desired results. Results reveal that proposed approach is suitable for the classification of the twitter messages. The limitation of the approach is different classifiers like ANN, SVM etc can be used to enhance classification performance. Bobicev [21] describes the classification of the sentiments in the textual form using machine learning. Actively developing area in the field of text mining is the analysis of subjects. Researchers explored that several classification task when texts are labelled by several sentiment labels and by this way the average F-measure reaches 0.805. Glinka et al. [22] states that in the information retrieval field multi-label text classification plays s significant role. It also discuss about the application of the feature extraction to develop effectiveness in the multi-label classification. The adequacy of the procedures is particularly essential on account of medical reports. They check the execution of the considered strategies by tests directed on the dataset of free medical textual reports. The results of the approach was compared with the two parameters i.e. classification accuracy and hamming gap .The limitation of the approach is normalization concept is not used. Bidi et al. [23] reviewed a feature selection methods based on genetic algorithms for different text representation methods. This states the empirical study of the component determination for the genetic algorithms for the different methods of text representation. This component determination calculation can achieve two objectives: in one hand is the inquiry of an element subset to such an extent that the execution of classifier is ideal. Secondly, is to discover an element subset with the smaller dimensionality that gives higher accuracy in classification. To assess the execution of this approach, three from the best classifiers have been chosen: Naive Bayes (NB), KNN and SVM. The goal is to decide if the genetic algorithms based on the component choice will enhance the exhibitions in content grouping with the smaller size using F-measure. The approach was tested on the two benchmarks 20-Newsgroup and Reuters-21578. Results shows that the SVM outperforms than other classifiers . In [24], two algorithms were used ABC algorithm was used to select the optimal features as well as SVM is used for the classification of the text and the proposed approach was tested on the UCI database. The limitation of the approach is that as only SVM is used different classifiers like ANN, NBC etc for parameters enhancement [25] states that the modern data consist of the larger amount of information which also contains the noisy data. It is significant to find the relevant features from the given raw data so the feature extraction as well as feature selection are the important concepts in analysing the data. It is with lower dimensionality the cost of the computing decreases. The lower dimensionality also decreases the complexity of data .

3. Proposed methodology

In our approach we have improved PCA for the extraction of features. In improved PCA normalization of extracted features are done with dividing by no of features observed. Mathematically, it can be shown as:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} X(i,j)}{r(x)*c(x)} \quad (1)$$

Where X denotes the data or feature value matrix after subtracting from mean, i=1,2,3…n,j=1,2,3,….m and r(x)and c(x) denotes the size of rows and columns of the X. It is normalization by size of features as it is easy to calculate as normalization with standard deviation has complex calculation. Further ABC is used for selecting features and svm for classification of texture.

3.1. PCA (principle component analysis)

It is a dimensionality reduction technique. It is important technique to remove irrelevant or redundant features [26], [27]. PCA plays a crucial role in various fields like: Artificial intelligence, biometrics etc [27]. Mathematically it can be stated as

When n observations of object Y are given in the d-dimensional space:

$$Y = \{y1, y2 \ldots yn\} \sum_{k=1}^{N} yk = 0, yk \in Rd \quad (2)$$

PCA computes the covariance matrix G by:

$$G = \frac{1}{N}\sum_{k=1}^{N} ykyk^{T} \quad (3)$$

to solve the equation for the eigen values and vectors

$$\lambda v = Gv \longleftrightarrow \lambda(yk.v) = (yk.Gv),\ k=1,2,N \quad (4)$$

Where
λ,v denotes Eigen vectors and values.
Whereas, in improved PCA we have added normalization with number of features calculation for feature reduction.
Algorithm of Improved pca :

1) Input the feature vector.
2) Subtract the mean data divided by size of features.
3) Perform the svd(single value decomposition)..
4) Evaluate the eigenvectors.
5) iterate i=1: size(Eigenvectors,1)
6) iterate j=1: size(Eigenvectors,2)
7) if(Eigenvectors(i,j) is greater than mean(Eigenvectors)
8) Eigenvectors(i,j) is equal to Eigenvectors(i,j)
9) else Eigenvectors(i,j) is equal to mean(Eigenvectors)

3.2. Artificial bee colony (ABC)

ABC is being used as an optimization technique [28]. It works with the three different bees: onlooker, employed and scouts Onlooker bees waits in the dance area as it gets the food source it turns into an employed bees. When employed bees consumes the food it becomes scouts. The amount of the nectar at the location of the food source .This value can be calculated as:

$$fiti = \frac{1}{1+fi} \quad (5)$$

Where artificial onlooker bee choose the food source through

$$pi = \frac{fiti}{\sum_{n=1}^{SN} fiti} \quad (6)$$

Where fit is the fitness value of the given solution i
Where SN is the no of the food source location
Algorithm of ABC

1) Initialize the selection process where a population of SN food source positions (i.e., applicant solutions) a_i, for i = 1, 2, SN. Each a_i is a vector of D parameters metrics : $a_i = [a_{i1}, a_{i2}, \ldots, a_{iD}]$ T
2) Compute the fitness of individual food source positions
3) repeat
4) Disturb individual food source positions ai to generate novel position zi
5) Calculate each new solution zi. If vi is enhanced than ai, then instead of ai use zi
6) Compute the probability pi related with each food source position ai.
7) For each processed bee, allocate it to a food source ai, that is proportionally based on the probability pi
8) Create new food source position zi by upsetting the food source ai of the individual processed bee.
9) Compute each novel solution zi. If zi is enhanced than ai, then replace ai with zi
10) If the food source is not enhanced during the previous limited cycles, then dump it and restore it with a new randomly positioned scout bees with its provisions source ai shaped by (4).
11) Learn the best food source positions that are found
12) Set series counter B = B + 1
13) until B = Maximum cycle number (MCN)
14) Return food source location (i.e., applicant solution) that is best establish so far.

3.3. (SVM support vector machine)

SVM is widely being used in the areas of classification and regression [29]. It is basically a binary classifier that can classify two classes. It separated the documents by drawing the hyper plane. Mathematically, the SVM can be stated as: Object Y has n dimensions Y=(y1,y2,…..yn) where $Yi \in R$ for i=1,2….n Each Y_j belongs to the class $Z_j \in \{-1, +1\}$. Training set T with the m pattern can be given by

$$T = \{(y1, z1)(y2, z2) \ldots (ym, zm\} \quad (7)$$

P is the space in the patterns Y are being embedded $y1, y2, y3 \ldots ym \in P$.Hyperplane in the space P, mathematically given as:

$$\{Y \in P | u.Y + t = 0\} u, t \in R \quad (8)$$

Dot product is given by

$$u.Y = \sum_{i=1}^{n} uiyi \quad (9)$$

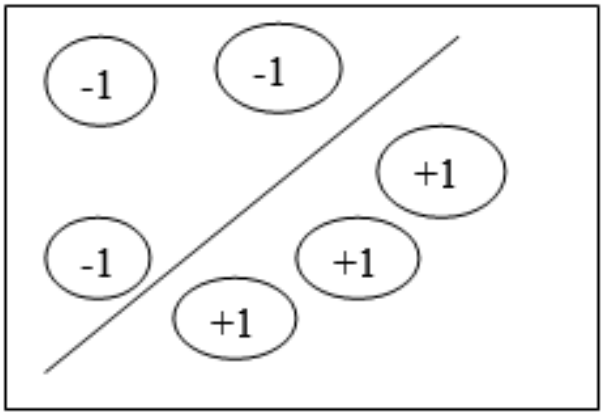

Fig. 1: SVMClassification.

This linear classifier is shown by the hyperactive plane $(u \cdot Y + t = 0)$ and divides or separates a region for class +1 patterns $(u \cdot Y + t > 0)$ and class −1 patterns is given by$(u \cdot Y + t < 0)$.
Algorithm of SVM

1) candidateVector = { nearest pair from the dissimilar classes }
2) while violating points are found do
   Search violator
3) candidateVector = candidateVector ∪ violator
4) if any ap< 0 with the addition of c to S then
5) candidateVector = candidateVector \ p
6) iterate till all such points are removed
7) end if
8) end while

Following flowchart represents the working of the proposed system:

1. Input raw data (noisy data)
2. Preprocessing is done (stop words removal and tokenization)
3. Features are extracted and normalized by PCA
4. Feature selection (ABC algorithm is used).
5. Finally the text classification (that is done through SVM)
6. Parameters are evaluated to check the efficiency of the proposed approach

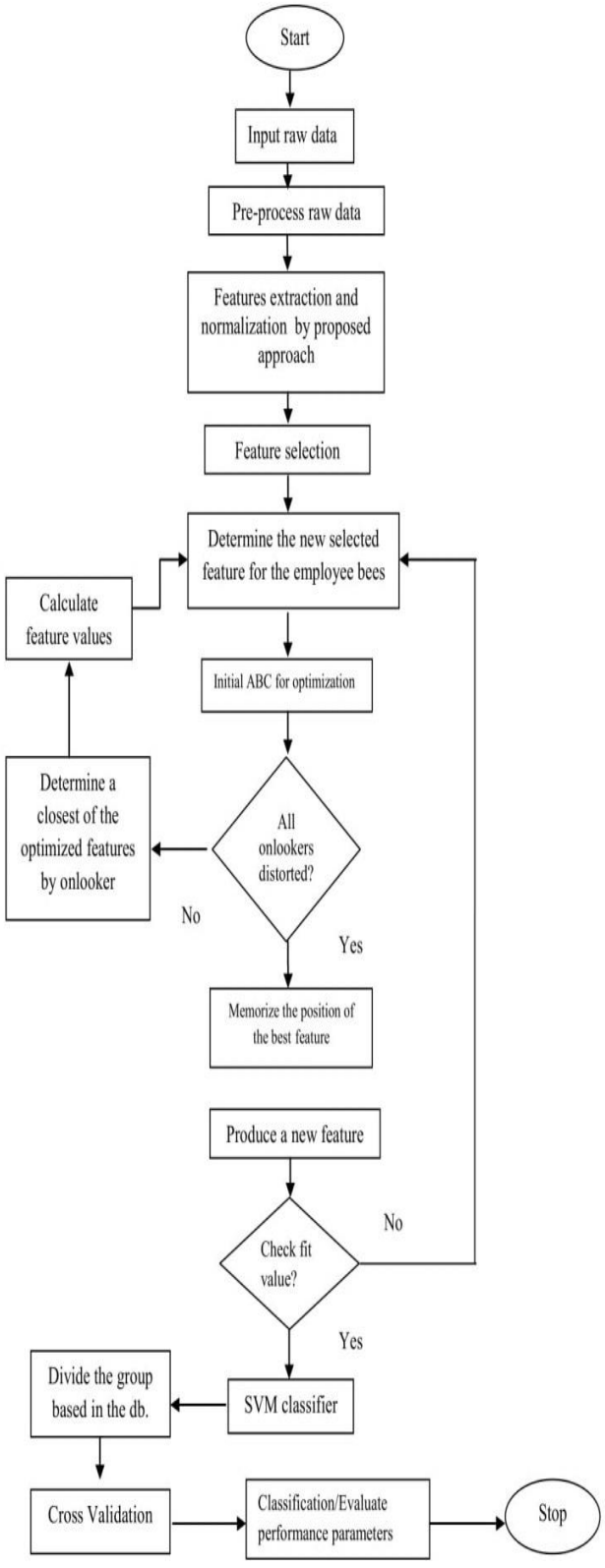

Fig. 2: Proposed Flowchart.

4. Experimental analysis

4.1. Dataset description

This section describes the dataset that is used in the proposed approach:
We use the 20-newsgroup dataset to test our proposed approach. It is a UCI standard dataset. It consists of the 20 different news categories. It is in the text format. It consists of large number of files in various categories. The different categories are about the atheism, graphics, sports etc. Each of the category contains approx 1000 files. This dataset is used also in many other researchers in their research works.

Table 1: 20-Newsgroups Dataset

Dataset description	
Name	20-Newsgroup
Type of data	Text format
Categories	20
Files in each category	1000approx

4.2. Result analysis

As the result of the proposed methodology. We get the better accuracy and f- measure

Table 2: Proposed Approach Result

Parameters	Proposed algorithm
Accuracy	97
MSE	0.40
F-measure	0.95
G-mean	0.84

Accuracy, F-measure, G-mean, MSE is mathematically calculated as:

Accuracy= abs(100*ROCAUC)(10)

Where ROCAUC=area under curve of true positive rate and false positive rate.

$F - measure = \frac{2*TPR*TNR}{TPR+TNR}(11)$

G-mean =$\sqrt{TPR * TNR}(12)$

Where TPR is the true positive rate and TNR is true negative rate. Accuracy rate value is 97, F-measure value is 0.95 and G-mean value is 0.84.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(Yi - yi)2 \quad (13)$$

Where Yi denotes the m no. of prediction values yi denotes the m no. of the actual values Table 3 shows the comparison of proposed and existing approach [11].

Table 3: Proposed Approach and Existing Approach Difference

Parameters	Proposed approach	Existing approach
F-measure	0.95	0.88
Accuracy	97	95

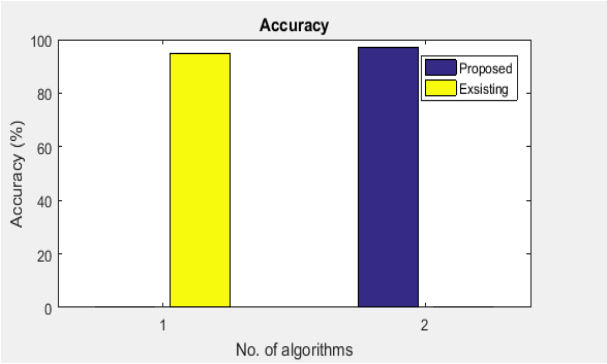

Fig. 3: Accuracy Analysis.

Figure 3 clearly defines the comparison between the accuracy of the existing approach and proposed approach. It can be analyzed that the proposed approach gives better accuracy.

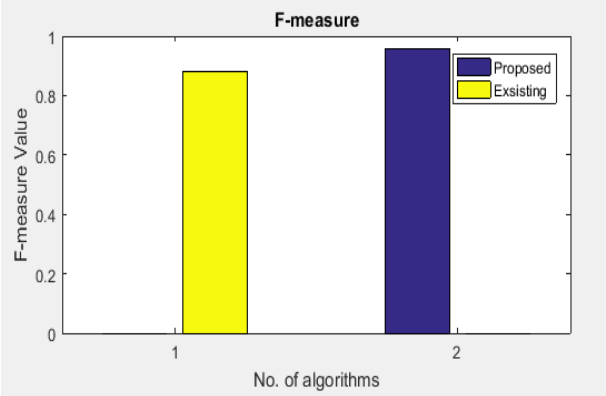

Fig. 4: F-Measure Analysis.

From the results it is observed that the proposed approach it provides the better accuracy and F-measure. It clearly defines the efficiency of the proposed approach it also enhances the performance as it give minimum mean squared error.

5. Conclusion and future scope

In this research work, we have used binary pre-processing methods namely stop word remove and tokenization on 20 Newsgroups. From the results, it can be seen that the pre-processing has a large impact on the performance of optimization classification. In our research work, we use the combination of the PCA, ABC and SVM for the classification of the text. Also, we tried to improve PCA through normalizing PCA, dividing by size of features .Performance parameters like accuracy value, F-measure, G-mean and MSE are calculated to analyze the performance of classifier .Experiment analysis proves that our proposed approach classify text with better accuracy. Only PCA, ABC and SVM is used in our research works different classifiers like Naïve Bayes, ANN, KNN etc can be used for the classification and accuracy values and different parameters can be analyzed Secondly, as our approach works on only small scale of data in future, it can be improvised to work on the large scale of data.

References

[1] S.A.Salloum, M.A.Emran, A.A.Monem, &K.Shaalen(2017) "Using Text Mining Techniques for Extracting Information from Research Articles",Intelligent Natural Language Processing: Trends and Applications,Vol.740,pp:373-397, Springer.

[2] B Jyot& G. Bathla (2018)," Document classification using various classification algorithms: a survey",,International journal of future revolution in computer science and communication engineering,vol.4,pp.150-155.

[3] P. L. Prasanna, D. R. Rao, Y. Meghana, K. Maithri& T. Dhinesh (2018),"Analysis of supervised classification techniques: International Journal of Engineering and technology, vol.7, pp.283-285, SPC.

[4] P.L.Prasanna&D.R.Rao (2018)"Text classification using artificial neural networks" International Journal of Engineering and technology, vol.7, no.1.1, pp.603-606, SPC.

[5] M.P Mali & M. Atique(2014) "Applications of Text Classification using Text Mining", International Journal of Engineering Trends and Technology (IJETT),Vol.13, no.5,SPC.

[6] J. Deepika, T. Senthil, C. Rajan& A. Surendar(2018),"Machine learning algorithms: a background artifact", International Journal of Engineering and technology,vol.7, pp.143-149,SPC.

[7] R.Thiyagarajana, S.Arulselvia& G. Sainarayanan (2010)," Gabor Feature based Classification using Statistical Models for Face Recognition", in Proceedings ofICEBT2 pp:83-93, Elsevier.

[8] A. Jain, K. Nandakumar& A. Ross (2005)'Score normalization in multimodal biometric systems", Pattern Recognition vol.38, pp. 2270 – 2285, Elsevier.

[9] D. Karaboga& B. Basturk,(2008) "On the performance of artificial bee colony (ABC) algorithm".Applied soft computing, vol .8, no.1, and pp: 687-697, Elsevier.

[10] C J.C.Burges& B. Schölkopf(1997), "Improving the accuracy and speed of support vector machines". In Advances in neural information processing systems, pp. 375-381.

[11] S. Zobeidi, M. Naderan& S. E. Alavi, (2017) "Effective text classification using multi-level fuzzy neural network", in proceedings of the 5th Iranian Joint Congress onFuzzy and Intelligent Systems (CFIS), pp. 91-96, IEEE.

[12] B.Tang, H. He, P.M.Baggantoss&S.kay (2016)"A Bayesian classification approach using class-specific features for text categorization." IEEE Transactions on Knowledge and Data Engineering vol.28, no.6 pp: 1602-1606.

[13] F.P.Shah&V.Patel(2016) ,"A review on feature extraction and Feature selection for text classification", Wispnet pp.2264-2268,IEEE.

[14] V. K. Vijayan, K. R. Bindu&L.Parameswaran(2017), "A comprehensive study of text classification algorithms." IEEE Advances in Computing, Communications and Informatics (ICACCI), pp: 1109-1113.

[15] M...S. Uzer, N. Yilmaz, & O. Inan (2013), "Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification" the scientific world journal,pp.1-10,Hindawi.

[16] Santoso, E. M. Yuniarno, &M.Hariadi (2015),"Large Scale Text Classification Using Map Reduce and Naive Baye's Algorithm for Domain Specified Ontology Building", in Proceedings of the 7th International Conference onIntelligent Human-Machine Systems and Cybernetics (IHMSC),vol. 1, pp. 428-432, IEEE.

[17] Y. Xue, J. Jiang, B. Zhao &T.Ma (2017)," A self-adaptive artificial bee colony algorithm based on global best for global optimization", Soft Computing, pp:1-18,Springer.

[18] M. Somvanshi,& P. Chavan (2016). "A review of machine learning techniques using decision tree and support vector machine", in Proceedings of the International Conference onComputing Communication Control and automation (ICCUBEA), pp. 1-7. IEEE, 2016.

[19] L. Demidova & I. Klyueva (2017)," SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem" in Proceedings of the sixth Mediterranean Conference on Embedded Computing (MECO) pp. 1-4. IEEE.

[20] K.Y. Wu, M. Zhou, X.S.Lu &L .huang (2017) "A fuzzy logic-based text classification method for social media data"in Proceedings of the International Conference on Systems, Man, and Cybernetics (SMC),, vol.13,no.3 pp:1942-19472,IEEE.

[21] V. Bobicev (2016). "Text classification: the case of multiple labels",in Proceedings of the International Conference on Communications (COMM) pp. 39-42. IEEE,

[22] K. Glinka, R. Woźniak, & D. Zakrzewska(2017), "Improving Multi-Label Medical Text Classification by Feature Selection",in Proceedings of the26th International Conference onEnabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) pp. 176-181. IEEE.

[23] N. Bidi, &Z.Elberrichi (2016),"Feature selection for text classification using genetic algorithms", in Proceedings of the 8th International Conference onModelling, Identification and Control (ICMIC), pp. 806-810. IEEE.

[24] H.wang, H.yu, Q.Zhang, S.Cang&W.Liao (2017)."Parameters optimization of classifier and feature selection based on improved artificial bee colony algorithm", in Proceedings of the International Conference on Advanced Mechatronic Systems (ICAMechS), IEEE.

[25] K.Modarresi(2015), "Unsupervised Feature Extraction Using Singular Value Decomposition", in Proceedings of the International Conference On Computational Science.vol.51,pp:2417–2425.Elsevier.

[26] H. Abdi, & L. J. Williams (2010)," Principal component analysis", Wiley interdisciplinary reviews: computational statistics, two, pp.1-47.

[27] T Meenpal, A.Goyal& A. Meenpal(2018),"Facial recognition system based on principle component analysis and distance measures ", International Journal of Engineering and technology,vol.7, no.2.21,pp.15-19,SPC.

[28] D. Karaboga& B. Basturk (2007)." A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm", Journal of global optimization, vol: 39 no.3, pp.459-471, Springer.

[29] A. J. Smola& B. Schölkopf,(2004), "A tutorial on support vector regression", Statistics and computing, vol.14, no.3, pp.199-222.