Effective Text Comment Classification Using Novel ML Algorithm - Modified Lazy Random Forest

Tejashri Ghodke and Vijay Khadse

August 2, 2020

Effective text comment classification using Novel ML algorithm - Modified Lazy Random Forest

Tejashri Ghodke
Department of Computer Engineering and IT
College of Engineering Pune, India
Email: ghodketa17.comp@coep.ac.in

V. M. Khadse
Department of Computer Engineering and IT
College of Engineering Pune, India
Email: vmk.comp@coep.ac.in

***Abstract*—Machine learning (ML) algorithms are methods used to classify data. The various patterns or classes can be classified with the help of these various ML algorithms. There are numerous areas where these algorithms can be used. One such area is to detect whether the comment, sms or text message is SPAM or Normal message. So the aim of this work is to identify the best machine learning algorithms to detect SPAM text message on two different dataset. The first dataset is collected from YouTube comment dataset and second is the SMS dataset. The Random Forest (RF) is the ensemble learning method for classification, regression and other tasks that operates by developing a multitude of decision trees at learning phase and outputting the class. Its one variant which performs well as compare to normal RF is Lazy RF, as the study shown and is the base for this research work. In this work, we have proposed one more novel variant of LRF and the different machine learning algorithms are compared in terms of accuracy with proposed Modified Lazy Random Forest. The results are compare with two techniques, first is the simple hold out, and second is the K-fold cross validation. For both cases the proposed algorithm performs well to detect the SPAM messages for the both datasets.**

***Index Terms*—IDS, Machine Learning, Deep Learning, Performance Matrices**

I. Introduction

The machine learning algorithms are used in various areas such as classification problems, regression problems, Virtual Personal Assistants, Traffic Predictions, Videos Surveillance and much more. One such area is SPAM message detection. The hard task was to find the best ML algorithm for detection of SPAM in text dataset. Then achieving the improvement in respective algorithm. The RF is the best choice in terms of promising results which it produces. The thorough literature review showed that LRF is well performing algorithm and we proposed the novel MLRF algorith. The comparative study showed that the proposed algorithm performs well as compared to current state-of-art algorithms. To cross verification we have found the results with hold out and K-Fold cross validation techniques. With hold out we have found out the results with different learning and training splits of the dataset. The L-RF and NB algorithms are analyzed on the YouTube API dataset. For the detailed insights the training and testing dataset split is considered from 10% Train - 90% Test, 20% Train - 80% Test, 30% Train - 70% Test, 40% Train - 60% Test, 50% Train - 50% Test, 60% Train - 40% Test and 70% Train - 30% Test. The completed work is performed using python programming language with tensorflow, numpy etc. libraries. The dataset pre-processing is done with the python with the built in hash function on string data. The processed dataset is then applied as an input to Modified Lazy-Random Forest and then to Nave Bayes algorithm. The Modified Lazy-Random Forest is performing well on the given dataset as compared to other compared algorithms.

II. Literature Review

The work done by N. Aggarwal et.al. in [1] has shown the classifier algorithmic rule for YouTube videos. The classifier is developed and tested with the sample dataset created by them. The YouTube contains tremendous variety of videos of assorted sorts as well as copyright profaned videos, business spam, hate and political theory promoting videos, vulgar and pornographic material and privacy incursive content.The obtained results established that accuracy of projected approach is quite 80%. They first type classifier approach to sight the privacy incursive harassment and offense videos having unwanted content on YouTube.. The analysis results found that the accuracy of those classifiers i.e. VVD, VAVDS, VAVDP and RVDC is 83%, 84%, 90%, and 97% severally. It indicates that known discriminatory options will be wont to exploit the harassment detection on YouTube unto an affordable accuracy.

The unstructured dataset creates problems in normal functioning of the popular machine learning algorithms in terms of classification, recognition and detection problems. The authors S. Phakhawat et.al. in [2] performed experiment on YouTube dataset and balanced the usage of the SMOTE technique along with examination of the usage of ubiquitous algorithms such as multinomial Nave Bayes (MNB), Decision tree (DT) and Support vector machines (SVM). After the experimentation SVM indicates promised results with an accuracy 93.30% on filtering task along with this 89.44% on classification problem. Their SMOTE approach could overcome the imbalanced data problem and provides an promised outcomes. Moreover, analyzing the results of SVM, the use of SMOTE provides an accuracy with 93.30% in considering to 76.41% with no re-sampling approach. The final result improved with 16.9% on Emotion Filtering dataset.

Video coding tutorials allow expert and novice programmers to visually examine actual builders write, debug, and execute code. Previous lookup in this area has focused on supporting programmers discover applicable content in coding tutorial movies as well as understanding the motivation and desires of content material creators [3]. A dataset of 6000 feedback sampled from 12 YouTube coding movies is used to behavior our analysis. The consequences also show that an extractive frequency-based summarization method with redundancy control, can sufficiently capture the important issues existing in viewers comments.

The researchers in [4] strived on comparative findings of the usual filtering procedures used for YouTube comment spams is carried. The study extended datasets obtained from YouTube, the use of its Data API. According to the retrieved results, excessive filtering accuracy (more than 98%) can be achieved with low-complexity algorithms, mentioning the possibility of acquiring a suitable browser extension to alleviate comment spam on YouTube. The expansion in social media reputation is crucial and should be a positive impact with greater apparent in the final decade. This extends in recognition as well as a need to motivated malicious, scammers, and spammers to goal these platforms.

The research work done in paper [5], by –S. Thiago et.al proposed a lazy model of the Random Forest (RF) classifier (named as LazyNN RF), primarily created for a noisy classification operations. The LazyNN RF localized coaching projection is composed via examples that higher resemble the examples to be classified, received via nearest neighborhood learning set projection. This gives marked evidence in favor of the exploring records regional in RF models. As future work, they deliberate to investigate distance metric learning.The theoretical analyses regarding the bias/variance trade-off of their proposal is also done.

The proposed model by A. Shreyas et.al. [6] for the detection of spam comments on the video-sharing website - YouTube is done with message classification as in two types, the promotional purpose and irrelevant. The authors strived to classify these comments by utilizing ML algorithms such as RF, SVM, Naive Bayes with specific custom heuristics such as N-Grams are proven to be very efficient in detecting spam comments. The authors have presented a method for automated detection of spam comments on the YouTube platform.

The research worked completed by done by A. Tripathy et.al. [7] used the IMDb movie review dataset. The primary goal of this paper classification of reviews on social platform websites into meaningful classes. For this purpose, they used many supervised ML algorithms such as NB, Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and SVM. Different machine learning algorithms are proposed for the classification of movie reviews of IMDb dataset (IMDb, 2011) using n-gram techniques viz., Unigram, Bigram, Trigram, a combination of unigram and bigram, bigram and trigram, and unigram and bigram and trigram. The proposed model gave 86.23% accuracy using Unigram+Bigram+Trigram (U+B+T) method for NB algorithm. The ME, SVM, and SGD gave accuracy 83.36%,88.94%,83.36% respectively using U+B+T method. The future research includes smiley comments which they not worked on.

The P. Sethi, V. Bhandary et.al. have worked on emails and messages with legal, economic and technical issues [8]. A pin point is being pone with the help pf Bayesian filters for preventing the SPAM messages and emails issue . They also analyzed and studied the relative strengths of various ML algorithms in for the detection of spam messages which are communicated on mobile devices. They collected the data from on open public dataset and developed two datasets for their testing and validation cases. Accuracy in detecting spam messages is considered as the base point for the comparing the results. Our results clearly demonstrate that various ML algorithms with different features leads to work differently in detecting spam messages

The work done in [9] by P. Kolari et.al. mainly focuses on Emails and Communication. The huge usage of emails in digital world brought many problems such as SPAM mail or Normal Mail. Their research focused on email content analysis in Turkish language. For classification purpose they used RF classifier algorithm and Vector Space Model from ML methods. Both techniques are subjected to different performance matrices and their performances are compared.

In the proposed system[10] by C. Chen et.al., a huge number of URL are scanned and analyzed with the help of many APIs for the identification of whether these URLs are malicious or not on time. They firstly carried the Spam Drift problem in statistical features based on Twitter spam detection. For solving this problem, They proposed a Lfun approach. Using this scheme, classifiers have been re-trained by the modified changed spam tweets that are learned from unlabelled patterns, which reduced the impact of Spam Drift significantly. They evaluated the performance of Lfun approach using performance matrices such as Detection Rate and F-measure. Their research concluded that both detection rate and F-measure are gained much when applying with their Lfun approach. They also compared Lfun to four traditional machine learning algorithms, and find that our Lfun outperforms all four algorithms in terms of overall accuracy, F-measure, and Detection Rate.

The email classification in two classes SPAM or Non-SPAM is carried by authors S. Olatunji[11] , with the help of Extreme Learning machines(ELM) and SVM for classification. In this work, they attempted investigation on how SVM and ELM compared to the unique and important problem of Email spam detection, which is a classification problem. Empirical results from experiments conducted using a ubiquitous dataset resulted that both techniques outperformed than the best earlier published techniques on the same popular dataset employed in this study. However, SVM performed better than ELM on a comparison scale based on accuracy. But in terms of speed of operation, ELM outperformed SVM significantly.

The authors V. Vishaghini et.al. [12] proposed the use of weighted SVM for spam classification using weight variables captured by the proposed KFCM algorithm. For the experiment UCI Repository, SMS Spam dataset is used with 4601 records having 57 attributes. Their research achieved a result that shows that WSVM with KFCM exhibits lower misclassification rate than SVM. In the end, they also analyzed the results on performance matrices such as accuracy and Precision. Their improved WSVM(with KFCM) achieved 96.5% accuracy.

III. System Design

A. Data pre-processing

Dataset:
The Datasets are YouTube SPAM dataset and SMS SPAM dataset. In which various messages and comments are present which are latest in nature. For the classification purpose we have pre-processed this dataset.

Numericalization:
The dataset consists of string values for some of the fields such as multilanguage pattern,non numeric values, therefore these attribute values are converted into numeric values to play better on them. Python language plays vital role here, with hash value function. After converting the all values into numeric values the dataset is given to training model.

B. Training with the Model

1) Modified Lazy Random Forest: It is obvious that the MLRF model consists of two parts - KNN processed dataset feature values for records and the second Random Forest part. Basically RNN gives processed data to Random Forest and then Random Forest gives detection capability. Due the the introduction of KNN the Random Forest performs lately and takes more time to generate results which make it named as Lazy Random Forest and we have modified the Lazy RF and proposed the MLRF.

2) Training set: Providing training dataset. With 50% training and 50% test dataset. Further experimentation is done with 55% Train- 45% Test subsequent up to 95%Train-5%Test. This technique of train and test is called "Hold-Out" method. K-Fold Croas validation technique is also applied for greater learning experience of the machine learning algorithm. The reason behind using this split for training and testing dataset is that it shows best accuracy and best performance in ML algorithms.

C. SPAM Classification

The last step of the architecture is the detection which classifies the record SPAM or Normal message. The proposed algorithm classifies both binary and multi-class classification. It is also possible to use this proposed algorithm anywhere where classification problem exist.

IV. Proposed Algorithm

The proposed algorithm is shown in following fig. 2. The algorithm is self explanatory and starts with the input dataset. The input dataset is then processed with KNN to identify the nearest neighbor and its distance, those values are then appended to base dataset record. The new processed dataset is then applied to Random Forest algorithm record by record i.e. line by line. This act caused the greater accuracy than simple Random Forest algorithm.

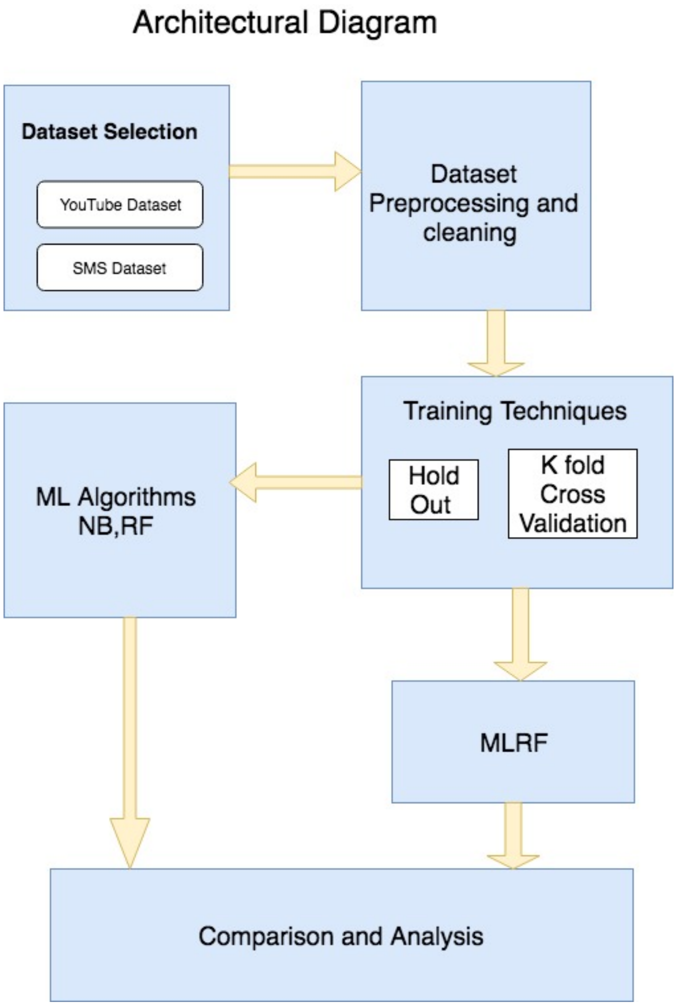


Fig. 1. System Architecture

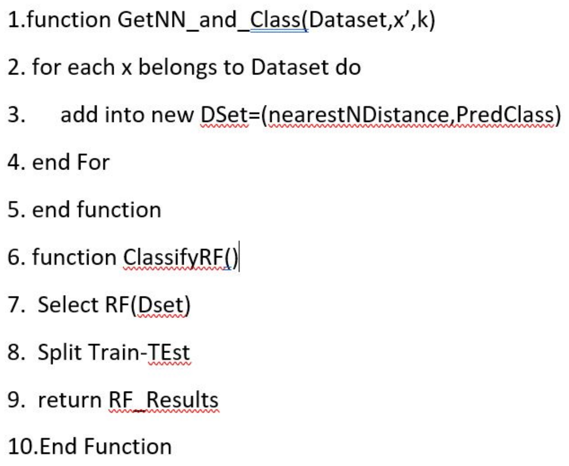
V. Experiment and Results

This section describes the experiment that we have done and results which we have obtained. The complete experiment is done in python programming language and various libraries are used such as scikit-learn, numpy etc. The experiment is done using hold out technique and the results are shown in table I.

VI. Results with Hold-Out method:YouTube and SMS Dataset

A. Hold-Out method results on YouTube Dataset

The first accuracy in above table I is for 50% Train-50%Test, second 55% Train- 45% Test subsequently last up to 95% Test-5% Test. The results are varying in nature because the program is taking the random records for its learning purpose. Same pattern of training- testing is spread across the following results. The same algorithms are applied on SMS dataset for the experimentation and the results are shown in



```
function GetNN_and_Class(Dataset,x',k)
for each x belongs to Dataset do
    add into new DSet=(nearestNDistance,PredClass)
end For
end function
function ClassifyRF()
Select RF(Dset)
Split Train-TEst
return RF_Results
End Function
```

Fig. 2. Proposed Algorithm

TABLE I
HOLD-OUT METHOD RESULTS ON YOUTUBE DATASET

Testing % Split	MLRF	RF	Nave Bayes
50	61.89175145	61.87175043	54.5445751221423
55	62.5352862	62.5049232	55.26593064
60	63.19597205	63.16359697	53.82534079
65	63.20912378	63.2	54.72081517
70	63.61755301	63.60259981	54.39428297
75	63.98035817	63.95147314	54.82993197
80	64.58103649	64.57204767	54.35992579
85	64.40829847	64.38949783	54.51109737
90	64.4198363	64.13095811	54.65883323

table 5.2 for SMS dataset with Hold out method. The results are compared and it is shown that the proposed Modified Lazy Random Forest is performing well over the compared algorithms.

B. Hold-Out method results on SMS Dataset

The results in table II clearly shows the MLRF gives promising results over other RF and Naive Bayes ML algorithm. The YouTube dataset contains 5 features and SMS dataset contains 2 features.

TABLE II
HOLD-OUT METHOD RESULTS ON SMS DATASET

Testing % Split	MLRF	RF	Nave Bayes
50	81.69657423	79.3800978793	55.90111643
55	81.42387077	78.88124439	56.52759085
60	80.70127002	79.4036444	58.17529472
65	81.43589744	79.20512821	57.83492823
70	81.00023929	78.84661402	57.28643216
75	81.48979134	78.23648194	38.02690583
80	81.48979134	78.39949324	56.69856459
85	82.34797297	75.88751496	58.60215054
90	81.75109693	78.1976195	51.61290323

TABLE III
5-FOLD CV METHOD RESULTS ON YOUTUBE DATASET

Logistic Regression	SVM	MLRF
0.505829596	0.802421525	0.807174888
0.507181329	0.836317774	0.822262118
0.517055655	0.800089767	0.833931777
0.488330341	0.830341113	0.816876122
0.509874327	0.802010772	0.829443447
Average	Average	Average
50.56542496	81.083619	82.17571713

TABLE IV
10-FOLD CV METHOD RESULTS ON SMS DATASET

LogisticRegression	SVM	MLRF
0.498207885	0.829749104	0.806451613
0.513464991	0.818671454	0.807899461
0.499102334	0.849192101	0.838420108
0.515260323	0.816624776	0.825852783
0.52064632	0.807648115	0.807899461
0.513464991	0.85432675	0.867145422
0.491921005	0.836624776	0.824057451
0.484739677	0.825852783	0.795332136
0.515260323	0.838420108	0.825852783
0.50448833	0.859964093	0.836624776
Average	Average	Average
50.5655618	82.07074059	82.5528143

VII. RESULTS WITH K-FOLD CV METHOD:YOUTUBE AND SMS DATASET

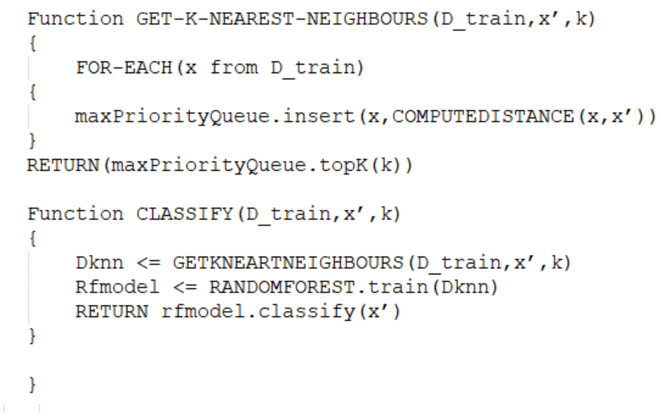
A. K-Fold CV method results on YouTube Dataset

The first accuracy in above fig table III is for YouTube dataset with first fold, second with accuracy of MLRF as 80.71% for second fold, third 82.22% for third fold and subsequently we have performed this task for five folds with K = 5 and the results are more interesting than hold out-method. The K-Fold cross validation has great learning capability for all the data records. The unseen data records are not covered in hold-out method but it is exactly apposite to K-Fold cross validation method. It learns all the data pattern records in the dataset for promising results.

The first accuracy in above table IV is for SMS dataset with first fold, second with accuracy of MLRF as 80.78% for second fold, third 83.84% for third fold and subsequently we have performed this task for five folds with K = 10 and the results are more interesting than hold out-method. The K-Fold cross validation has great learning capability for all the data records. The unseen data records are not covered in hold-out method but it is exactly apposite to K-Fold cross validation method. It learns all the data pattern records in the dataset for promising results. The same results are compared with the results of other ML algorithms i.e. logistic regression and SVM.

B. Lazy Random Forest

More formally, assume Dtrain = (xi , yi)is the learning set of records. Actual learning task is started when this learning records are received by classified. The nearest x are identified for the purpose of projecting the training set to a subset of



```
Function GET-K-NEAREST-NEIGHBOURS(D_train,x’,k)
{
    FOR-EACH(x from D_train)
{
    maxPriorityQueue.insert(x,COMPUTEDISTANCE(x,x’))
}
RETURN(maxPriorityQueue.topK(k))

Function CLASSIFY(D_train,x’,k)
{
    Dknn <= GETKNEARTNEIGHBOURS(D_train,x’,k)
    Rfmodel <= RANDOMFOREST.train(Dknn)
    RETURN rfmodel.classify(x’)
}

}
```

Fig. 3. LRF Algorithm

training records for finding the class of x. Assume that x is given, the set Dknn(x) contains k nearest neighbors of x which are computed and used for learning task of RF classifier. Many distance matrix techniques can be used here: (i) The inverse of cosine similarities, (ii) adaptive distance metrics that define non-isotropic neighborhood based on the observed characteristics of the input space, (iii) distance matrix learned from training records. They here considered the simple neighborhood definition based on cosine similarities. These procedures are executed for each test record for classification. The strategy is outlined in the algorithm below:

C. Naive Bayes

Naive Bayes(NB) is being studied thoroughly since the 1960s. It was introduced (even not given that name) into the text retrieval community in the 1960s, and persists a popular technique for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines.[2] It also finds application in automatic medical diagnosis. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in terms of the number of variables (features) in a learning tasks.

VIII. Conclusion

Initially the results for comment message detection of SPAM or normal is done with help of current state-of-art algorithms such as Random Forest, Naive Bayes, Lazy Random forest etc. and those results are compared with the proposed algorithm. The results are more promising than those current state-of-art algorithms, the results have proved that with accuracy, false positive rate, precision, recall and with the confusion matrix. To refine the experimentation and results in thoroughly, the two techniques have been implemented called hold out and K-Fold Cross validation. These two techniques have shown the promising results. The numericalization step has been introduced for values for the attributes. The proposed MLRF has best performing capacity on text comment classification. Moreover, the proposed algorithm can be applied to any of the classification problems such as in security field for attack detection, smart predictions and much more. The future work includes applying the various different dataset of different domain and then analyze how the proposed MLRF performs.

References

[1] Aggarwal, Nisha, Swati Agrawal, and Ashish Sureka. ”Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos.” In Privacy, Security and Trust (PST), 2014 Twelfth Annual International Conference on, pp. 84-93. IEEE, 2014.

[2] Sarakit, Phakhawat, Thanaruk Theeramunkong, and Choochart Haruechaiyasak. ”Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm.” In Advanced Informatics: Concepts, Theory and Applications (ICAICTA), 2015 2ndhttps://www.overleaf.com/project/5cc57357f0e26d3383f371fd International Conference on, pp. 1-5. IEEE, 2015.

[3] Poch, Elizabeth, Nishant Jha, Grant Williams, Jazmine Staten, Miles Vesper, and Anas Mahmoud. ”Analyzing user comments on YouTube coding tutorial videos.” In Proceedings of the 25th International Conference on Program Comprehension, pp. 196-206. IEEE Press, 2017.

[4] Abdullah, Abdullah O., Mashhood A. Ali, Murat Karabatak, and Abdulkadir Sengur. ”A comparative analysis of common YouTube comment spam filtering techniques.” In Digital Forensic and Security (ISDFS), 2018 6th International Symposium on, pp. 1-5. IEEE, 2018.

[5] Salles, Thiago, Marcos Gonalves, Victor Rodrigues, and Leonardo Rocha. ”Improving Random Forests by Neighborhood Projection for Effective Text Classification.” Information Systems (2018).

[6] Aiyar, Shreyas, and Nisha P. Shetty. ”N-Gram Assisted Youtube Spam Comment Detection.” Procedia Computer Science 132 (2018): 174-182.Abinash Tripathy, Ankit

[7] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, Classification of Sentiment Reviews using N-gram Machine Learning Approach, Expert Systems With Applications (2016), doi: 10.1016/j.eswa.2016.03.028

[8] Sethi, P., Bhandari, V. and Kohli, B., 2017, October. SMS spam detection and comparison of various machine learning algorithms. In 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN) (pp. 28-31). IEEE.

[9] Kolari, P., Java, A., Finin, T., Oates, T. and Joshi, A., 2006, July. Detecting spam blogs: A machine learning approach. In Proceedings of the national conference on artificial intelligence (Vol. 21, No. 2, p. 1351). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[10] Kamble, S. and Sangve, S.M., 2018, August. Real Time Detection of Drifted Twitter Spam Based on Statistical Features. In 2018 International Conference on Information, Communication, Engineering and Technology (ICICET) (pp. 1-3). IEEE.

[11] Olatunji, S.O., 2017, April. Extreme Learning machines and Support Vector Machines models for email spam detection. In 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE) (pp. 1-6). IEEE.

[12] Vishagini, V. and Rajan, A.K., 2018, August. An Improved Spam Detection Method with Weighted Support Vector Machine. In 2018 International Conference on Data Science and Engineering (ICDSE) (pp. 1-5). IEEE.