Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

Effective Adaboost Sequential Classification Algorithm based Ensemble Method for E-Mail Spam Filtering

1T.Poonkodi, 2Dr.S.Sukumaran
1Ph.D Research Scholar, 2Associate Professor
Department of Computer Science,
Erode Arts and Science College (Autonomous),
Erode, India.
1ponrohit.0707@gmail.com

Abstract: Email spam is still a problem even today, and spammers still approach it the spam way. Spam accounts for billions of emails sent everyday which makes up most of all emails. For email spam detection, Sine-Cosine Algorithm (SCA) was used to which enable the selection of optimal features and the feature vectors are updated. It returned the spam detection based on decreased feature selection error. Even so, the large collection of emails in the datasets makes it difficult to be faced with more storage space and spam detection time. An Enriched Firefly Optimization Algorithm (EFOA) was introduced selecting the best features and the weights of the features are updated using optimized semantic WordNet for email spam detection. The adapted, scalable and integrated filters with Effective AdaBoost Sequential Classification based Ensemble Method (EASCEM) for email spam detection is proposed in this paper. In EASCEM, the advanced differential grading weighting schemes used based on the classified error rate. The Effective AdaBoost sequential classifiers are used to classify the email text message as spam and non-spam. The experiments results are proved the effectiveness of the proposed EASCEM method.

Keywords: EFOA, AdaBoost Classification, Optimized semantic WordNet, Sine-Cosine Algorithm

I. Introduction

Spam has become a critical threat which reducing the use of electronic emails as a means of communication medium [1]. The continued rise of spam emails has led to mitigating approaches to protect e-mail users by filtering or rejecting them all together. The need for spam filtering which essentially cossists of separating spam from spam emails using computational tools. An even more important step in the filtering process is to detect whether an email is spam or not as that determines what is done at the back of the email [4].

Boosting is a general technique used to increase the accuracy of any given learning algorithm. The AdaBoost algorithm was initially defined for two class issues, but it can be further defined for multi-class and regression issues [2]. Stimulation approaches have been proposed as overall methods that depend on the principle of generating multiple predictions and a majority vote

among individual classifiers. A new phishing email detection model is used to model emails into the header, body of the email, character level and also word level concurrently [3].

In order to deal with spam detection error in email spam detection system, spam detection methodology [5] was projected by means that Sine-Cosine Algorithm. This methodology uses feature selection strategies and feature vectors are updated by the SCA for preparing the ANN that reduced the spam detection error. But, the matter of space and time complexities are increased in E-Mail spam detection is more difficult for giant collection of emails. So, Enriched Firefly Optimization Algorithm (EFOA) [6] was proposed to enhance and recovering the problem of space and time complexities. In EFOA, the optimized semantic WordNet is utilized for clean the noise data, semantic based feature reduction and have weights are updated. Then selecting suitable features using Enriched Firefly Optimization Algorithm (EFOA) and ANN classifier classifies the email as spam and non-spam which improved the performance of email spam detecting system.

In this paper, Effective AdaBoost Sequential Classification based Ensemble Method (EASCEM) is proposed for efficient email spam detection. In EASCEM, the sample weights are calculated depending on the email features and create the base learner model. Every email records gets the same weight. The total error of incorrectly classified mails and performance of stumps are calculated. The advanced differential grading weighting schemes are assigned based on incorrectly classified emails. All emails are classified as spam and non-spam using EASCEM Classifier.

II. Related Works

Yan et al., **[7]** proposed method for classifying multi-tag documents using the word victimation2 with LSTM and Connectionist Temporal Classification (CTC). This model is evaluated on totally different datasets together with e-mails and produce promising results compared to alternative versions of each sequent deep learning models like RNN and non-sequential algorithms like Support Vector Machines (SVM). Their research attempts to solve multi-label classification problems by first representing the document with an LSTM network, then forming another LSTM network to represent the classified label stream. Finally, they use the CTC to predict more than one label.

Sunday Olusanya Olatunji et al., [8] proposed Support Vector Machines (SVM)-based model is investigated toward achieving higher accuracy of spam detection whereas paying robustattention to befittingly victimization thoroughgoing parameter search techniques to make sure higher spam detection accuracy. The SVM has become the focus of several recent researchers because of its ability to generalize appropriately to either few data samples or huge data samples. SVM has its theoretical basis in supporting statistical theory by deploying its powerful phenomenon otherwise known as a core trick.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

Devottam Gaurav et al., [9] proposed a new method of detecting spam based on the concept of document labeling that classifies new spam and ham. Furthermore, algorithms such as Naive Bayes, Decision Tree and Random Forest (RF) are employed in the classification process. Three datasets are used to assess the functionality of the proposed algorithm. Experimental results show that RF has a greater accuracy compared to other methods. Due to limited convenience of datasets for email spam, constrained data and therefore the text written in an informal way are the foremostmost possible problems that forced the present algorithms to fail to meet the expectations throughout classification.

Nandan Parmar et al., [10] proposed a built-in approach using the Naïve Bayes algorithm as well as Particle Swarm Optimization is used for email spam detection. The naïve Bayes algorithm is used to learn and classify e-mail as spam and ham. Particle Swarm Optimization is a stochastic optimization technique and is used for worldwide heuristic optimization of Naive Bayes parameters. For experimentation, the Ling Spam dataset is considered and the outcome is evaluated in terms of accuracy, f-measurement, accuracy and recall.

Siti Aqilah Khamis et al., [11] proposed identifying potentially useful e-mail header functions for email spam detection by analyzing two sets of email data; challenges in detecting anomalies and exploring cybersecurity data from the Web site. By analyzing the datasets, the most objective of this analysis is to extract the appropriate options of the email header and examine the features options to classify the features using Support Vector Machine (SVM) victimisation RapidMiner Studio and Weka. The methodology is divided into five phases: data collection, data pre-processing, characteristic selection, classification and detection. Email header categorization using Support Vector Machine (SVM) for CSDM2010 is higher than Anomaly Detection Challenges datasets.

Hossam Farisa et al., [12] proposed an intelligent detection system based on Genetic Algorithm (GA) and Random Weight Network (RWN) to handle email spam detection tasks. Additionally, an automatic identification function is also incorporated into the proposed system to detect the most relevant features during the detection process. The proposed system is being intensively evaluated using a series of in-depth experiments based on three email collections. The experimental results confirm that the proposed system can provide outstanding results in terms of accuracy, accuracy and recall.

III. Proposed Methodology

The emails are collected from different datasets. Optimized semantic WordNet is reduced the extracted textual feature and updating weighted feature. The best feature is selected using EFOA. The selected features are fed into EASCEM. The EASCEM method classifies the emails

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

as spam and non-spam. It enhances the accuracy of email spam filtering [15]. Figure 1 shows the architecture of EASCEM.

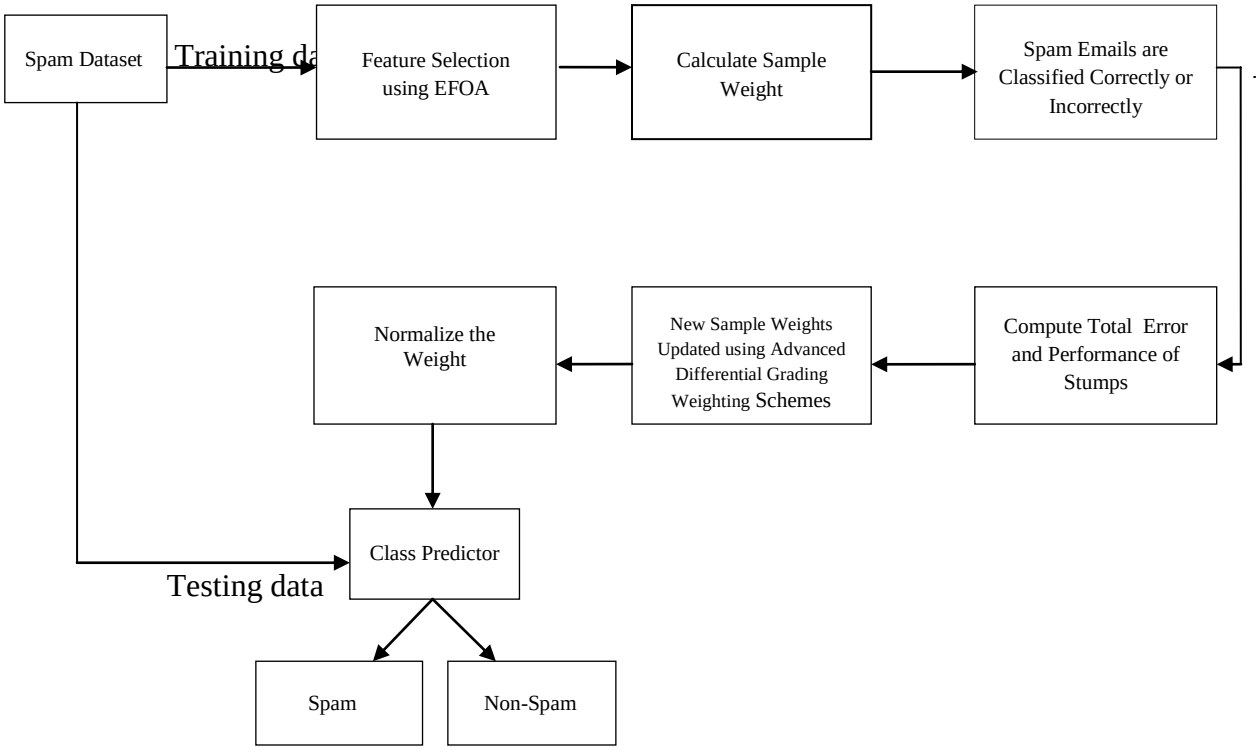
3.1 Calculation of sample weight

The term "boost" refers to a family of algorithms which converts weak learners into strong learners. Boosting is a machine learning approach based on the idea of creating a highly accurate prediction rule by combining numerous relatively weak and incorrect rules. Adaptive Boosting consists of reassessing data rather than random sampling. This method develops a system construction concept to improve the efficiency of classifiers. The AdaBoost (Adaptive boosting) classifier builds a strong classifier by combining a number of underperforming classifiers.

In AdaBoost, the goal is to combine different weak learners or classifiers ($h_i(x)$) together to enhance the classification performance, where $h_i(x)$ is a unique classifier. Every weak learner is trained with a simple set of training samples. Each sample has a weight and all samples have their weights are adjusted iteratively. AdaBoost iterately trains weak learners and calculates a weight for each of them and this weight represents the strength of the weak learner. All email records are provided with a sample weight. To assign sample weight, the formula used is,

$$\omega = \frac{1}{N} \qquad (3.1)$$

Where N is the number of records.



Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

Figure 1: Block Diagram of EASCEM

For choosing a base learner, there are two properties are Gini and Entropy. The stump that has the smallest value will be the primary base learner. The number of leaves represents the correctly and incorrectly classified records. By using these records, the Gini or entropy index is calculated. The stump that has the least entropy or Gini will be selected for the base learner. The final AdaBoost classifier is

$$H_T = sign(\sum_{t=1}^{T} \alpha_t h_t(x) \quad (3.2)$$

Where T is the total number of weak learners and x is a sample. Hence $H_T = \alpha_1 h_1(x) + \alpha_2 h_2(x) + \cdots + \alpha_T h_T(x)$ and $H_T(x) = H_{T-1}(x) + \alpha_t h_t(x)$. AdaBoost uses exponential loss functions that is outlined as follows

$$L\left(y_i, f(x_i)\right) = e^{-y_i . f(x)} = e^{-y_i . H_T(x_i)} \quad (3.3)$$

The total error is that the sum of all the errors within the classified record of sample weights. calculate error rate for every weak classifier is

$$\epsilon = \sum_{Wrong} \omega_i \quad (3.4)$$

In boosting, soley the incorrect emails or wrongly classified records got a lot of most popular than the correctly classified records. Thus, only the incorrect records from the stump are passed on to a different stump. While in AdaBoost, each records were allowed to pass, the incorrect records are recurrentquite the right ones.

Increase the weight of the wrongly classified records and reduce the weight for the properly classified records. By putting the value of total error in the above formula and when determination to get the value for the performance of stump.

$$\boldsymbol{\alpha} = \frac{1}{2}\log\frac{1-\epsilon}{\epsilon} \quad (3.5)$$

3.2 New Weights Updated using Advanced Differential Grading Weighting Scheme

Updating the weights based on the performance of the stump. Update the sample weight before continuing for ensuing model or stage as a result of it identical weight is applied, receive the output from the primary model. For incorrectly classified email records the formula is:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

$$\omega_{new} = \begin{cases} \frac{\omega_{old}}{2(1-\epsilon)} & if\ point\ classified\ correctly \\ \frac{\omega_{old}}{2\epsilon} & if\ point\ classified\ wrongly \end{cases} \quad (3.6)$$

In this classifier, as opposed to assigning identical vote to the weights of wrongly classified instances anytime. The advanced differential grading weighting scheme is outlined based on error rates**.** In the event that the error rate is between 0.1 and 0.2, the misclassified occurrences are given lesser votes as compared to the votes given in case the error rate is between 0.3 and 0.4.

The normalized weight is calculated as a division of sum of old weights and sum of latest weights.

$$N = \frac{\omega_{old}}{\omega_{new}} \quad (3.7)$$

In the new dataset, the frequency of incorrectly classified records are going to be quit the right ones. A new dataset created supported normalized weights. It select the incorrect records for training purposes which will be the second decision stump. To form a new dataset based on normalized weight, the calculation will partition it into buckets.

Predict the classified result using ensemble method of effective adaboost sequential classification. Initialize the weight of every class to zero. For each of the k classifiers, include weight to the class predicted by M(f).

$$W_i = \log\frac{1-Error\,M(f)}{Error\,M(f)} \quad (3.8)$$

Return the class having the biggest weight.

EASCEM Algorithm

Step 1: Initialize Spam Datasets.
Step 2: Create a decision stump for each emails.
Step 3: Calculate total error rate for weak classifier in classified emails using eq.

$$\epsilon = \sum_{Wrong} \omega_i$$

Step 4: Pick Classifier with the lowest error rate.
Step 5: Compute performance power for the classifier using eq.

$$\boldsymbol{\alpha} = \frac{1}{2}\log\frac{1-\epsilon}{\epsilon}$$

Step 6: Calculate error rate for classified and incorrectly classified emails using eq.

$$\omega = \begin{cases} \frac{\omega_{old}}{2(1-\epsilon)} & if\ point\ classified\ correctly \end{cases}$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

IV. Results and Discussion

The performance of proposed EASCEM is compared with ALO-Boosting based email spam detection method in terms of accuracy, precision, recall, False Positive Rate, False Negative Rate and time. For the experimental purpose, the emails are collected from spambase in UCI repository, enron spam publicly available dataset.

4.1 Evaluation Metrics

Accuracy

Accuracy is computed as the percentage of the dataset correctly categorized by the algorithm. The percentage of total number of properly recognized e-mails defined by the following formula:

$$Accuracy = \frac{No\ of\ correctly\ classified\ non\ spam\ emails + No\ of\ correctly\ classified\ spam\ emails}{Total\ No\ of\ spam\ emais + Total\ No\ non\ spam\ emails} \quad (4.1)$$

Precision

Precision indicates the number of jurisdictions which are positively ranked and relevant. High precision demonstrates the high pertinence for positive detection.

$$Precision = \frac{Correctly\ classified\ non\ spam\ emails}{Correctly\ classified\ non\ spam\ emails + Falsely\ classified\ non\ spam\ emails\ as\ spam} \quad (4.2)$$

Recall

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

Recall is defined as the probability of correctly classifying spam e-mails as spams, and the legitimate recall is defined as the probability of properly classifying correctly legitimate e-mails. The Recall formulas are listed below:

$$Recall = \frac{Correctly\ classified\ non\ spam\ emails}{Correctly\ classified\ non\ spam\ emails + Falsely\ classified\ spam\ emails\ as\ non\ spam} \quad (4.3)$$

False Positive Rate

The false positive rate defines the error in judgment ratio of legitimate messages as spam. The FPR formulas are listed below:

$$FPR = \frac{Falsely\ classified\ non\ spam\ emails\ as\ spam}{Falsely\ classified\ non\ spam\ emails\ as\ spam + correctly\ classified\ spam\ emails} \quad (4.4)$$

False Negative Rate

The false negative rate defines the spam mail misjudging ratio as legitimate. The FNR formulas are listed below:

$$FNR = \frac{Falsely\ classified\ spam\ emails\ as\ non\ spam}{Falsely\ classified\ spam\ emails\ as\ non\ spam + Correctly\ classified\ non\ spam\ emails} \quad (4.5)$$

Time

It measures the amount of time taken to filtering the spam emails from the database.

4.2 Performance Measures for Spam Dataset

The performance analysis of EASCEM and ALO-Boosting on spam dataset is shown in Table 1.

Table 1 Performance Measure for Spam Dataset

Methods	Precision (%)	Recall (%)	False Positive Rate	False Negative Rate	Time (Sec)
SCA	98.64	96.52	0.132	0.166	10.4
ALO-Boosting	98.90	97.01	0.060	0.101	7.7
EASCEM	99.21	97.95	0.039	0.071	6.2

The precision, Recall, False Positive Rate, False Negative Rate, time of EASCEM based email spam filtering method is greater than ALO-Boosting and SCA based email spam filtering method on spam dataset. From this analysis, it is proved that proposed EASCEM method has high precision, Recall, False Positive Rate, False Negative Rate, time than ALO-Boosting method on the spam dataset for email spam filtering.

4.3 Performance Measures for Enron Spam Dataset

The performance analysis of EASCEM and ALO-Boosting on Enron spam dataset is shown in Table 2.

Table 2 Performance Measure for Enron Spam Dataset

Methods	Precision (%)	Recall (%)	False Positive	False Negative	Time

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

			Rate	Rate	(Sec)
SCA	94.10	92.16	0.153	0.197	11.9
ALO-Boosting	97.22	95.31	0.080	0.145	9.1
EASCEM	98.64	96.33	0.058	0.99	7.5

The precision, Recall, False Positive Rate, False Negative Rate, time of EASCEM based email spam filtering method is greater than ALO-Boosting and SCA based email spam filtering method on enron spam dataset. From this analysis, it is proved that proposed EASCEM method has high precision, Recall, False Positive Rate, False Negative Rate, time than ALO-Boosting method on the enron spam dataset for email spam filtering.

4.4 Accuracy Comparison for Spam Dataset

The performance analysis of SCA, ALO-Boosting and EASCEM on Spam dataset is shown in Table 3.

Table 3 Filtering of Spam and Non Spam Accuracy Measures for Spam Dataset

Methods	Spam Accuracy(%)	Non Spam Accuracy(%)	Over all Accuracy(%)
SCA	97.51	98.33	97.92
ALO-Boosting	97.68	98.76	98.22
EASCEM	98.32	98.89	98.60

The spam and non-spam accuracy of EASCEM based email spam filtering method is greater than ALO-Boosting and SCA based email spam filtering method on spam dataset. From this analysis, it is proved that proposed EASCEM method has high accuracy than ALO-Boosting method on spam dataset for email spam filtering.

4.5 Accuracy Comparison for Enron Spam Dataset

The performance analysis of EASCEM and ALO-Boosting on the Enron spam dataset is shown in Table 4.

Table 4 Filtering of Spam and Non Spam Accuracy Measures for Enron Spam Dataset

Methods	Spam Accuracy(%)	Non Spam Accuracy(%)	Over all Accuracy(%)
SCA	92.39	93.97	93.18
ALO-Boosting	95.96	97.64	96.80
EASCEM	96.63	98.12	97.37

The spam and non-spam accuracy of EASCEM based email spam filtering method is greater than ALO-Boosting and SCA based email spam filtering method on enron spam dataset. From this analysis, it is proved that proposed EASCEM method has high accuracy than ALO-Boosting method on the Enron spam dataset for email spam filtering.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 13056 - 13066
Received 05 March 2021; Accepted 01 April 2021

V. Conclusion

This paper proposed an advanced differential grading weighting scheme for effective email detection system. The selected features are updating the sample weight and creating the base learner model. In this base learner model, the emails are classified as correctly and incorrectly. For incorrectly emails, the total error and performance of stumps are calculated. The advanced differential grading weighting scheme assigned for all incorrect emails. Based on the normalized weight, the new model is created. The emails are classified into spam and non- spam using EASCEM method. The experimental results shows that the proposed EASCEM has better accuracy, precision, recall, false positive rate, false negative rate and time than other email spam detection methods. Thus the proposed method is suitable for effective email spam detection.

References

[1] Huwaida T. Elshoush & Esraa A. Dinar, "Using Adaboost and Stochastic gradient descent (SGD) Algorithm with R and Orange Softwares for Filtering email spam", IEEE, Vol. 8, No. 9, pp. 41-46, 2019.

[2] Heba Gamal, Nour Eldin Ismail & Rizk M. R. M, "A Coherent Performance for Noncoherent Wireless Systems Using AdaBoost Technique", MDPI, Vol. 9, No. 2, pp. 1-9, 2019.

[3] Fang, Y., Zhang, C., Huang, C., Liu, L., Yang, Y.: Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism. IEEE Access, Vol. 7, pp. 56329–56340, 2019.

[4] Ismail B. Mustaphaa, Shafaatunnur Hasana & Sunday O. Olatunji, "Effective Email Spam Detection System using Extreme Gradient Boosting", SCI, pp. 1-10, 2020.

[5] Roztta Talael Pashiri, Yaser Rostami & Mohsen Mahrami, "Spam detection through feature selection using artificial neural network and sine-cosine algorithm", Springer, No. 14, pp. 193-199, 2020.

[6] Hamsapriya .T, Karthika Renuka .D & Raja Chakkaravarthi .M, "Spam classification based on supervised learning using machine learning techniques", MDPI, Vol. 2, No. 4, 2015.

[7] Yan .Y, Wang .Y, Gao .WC, Zhang .BW & Yin XC, "Lstm2 : multi-label ranking for document classification", Neural Process Lett., Vol. 47, No. 1, pp. 117-138, 2018.

[8] Sunday Olusanya Olatunji, "Improved email spam detection model based on support vector machines", Neural Comp. and Applic., Springer, 2017.

[9] Devottam Gaurav, Sanju Mishra Tiwari & Ayush Goyal, "Machine intelligence-based algorithms for spam filtering on document labeling", Soft Computing, Springer, 2019.

[10] Nandan Parmar , Ankita Sharma & Harshita Jain, "Email Spam Detection using Naïve Bayes and Particle Swarm Optimization", International Journal of Innovative Research in Technology, pp.367-374, Vol. 6, No. 10, 2020.

[11] Siti Aqilah Khamis, Cik Feresa Mohd & Nordiana Rahim, " Header Based Email Spam Detection Framework Using Support Vector Machine (SVM) Technique", Int. Conf. on Soft Computing, Springer, pp. 57-65, 2020.

[12] Hossam Farisa, Ala M. Al-Zoubia, Ali Asghar Heidarib & Ibrahim Alijaraha, "An Intelligent System for Spam Detection and Identification of the most relevant features based on evolutionary Random Weight Networks" , Elsevier, 2018.

[13] Tang, Ding. Z & Zhou .M, "A Spammer Identification Method for Class Imbalanced Weibo Datasets", IEEE Access, Vol. 7, pp. 29193-29201, 2019.

[14] Youwei Wang & Lizhou Feng, "Improved Adaboost Algorithm for Classification based on Noise Confidence Degree and Weighted Feature Selection", IEEE Access, Vol. 8, pp. 153011-153026, 2020.

[15] Tang.J, Deng.C & Huang.G, "Extreme learning machine for multilayer perceptron", IEEE Trans. NeuralNetw. Learn. Syst., Vol. 27, pp. 809–821, 2016.

[16] Doaa Mohammed Ablel-Rheem, Ashraf Osman Ibrahim & Shahreen Kasim, "Hybrid Feature Selection and Ensemble Learning Method for Spam Email Classification", Int. Jou. Of Adv. in Com. Sci Tech, pp. 217-223, Vol. 9, No. 14, 2020.

[17] Bassiouni .M, Ali .M & El-Dahshan, "Ham and Spam E-Mails Classification Using Machine Learning Techniques", Journal of Applied Security Research, Vol. 13, No. 3, pp. 315-331, 2018.

[18] Mohamad .M & Selamat .A, "An evaluation on the efficiency of hybrid feature selection in spam email classification", IEEE Communications and Control Technology (I4CT), pp. 227-231, 2015.

[19] Karthika Renuka .D, Visalakshiand .P, Rajamohana .SP, "An Ensembled Classifier for Email Spam Classification in Hadoop Environment", Appl. Math. Inf. Sci., Vol. 4, No. 11, pp. 1123-1128, 2017.

[20] Izzat Alsmadi & Ikdam Alhami, "Clustering and Classification of email contents", Journal of King Saud and Information Sciences, pp. 46-57, 2015.