International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Email Spam Detection using Ensemble Methods

Uma Bhardwaj, Priti Sharma

Abstract*: The swiftly growth of spam email has escalated the need to upgrade the existing spam detection and filtration methods. There is the existence of several machine learning methods for the classification and detection of email spam but these lacks in some cases. In this research work ensemble methods are adapted to detect the email spam. The machine learning methods of Multinomial Naïve Bayes and J48 Decision Tree algorithms are considered and ensembled. The considered ensemble methods are bagging and boosting. The experimentation is conducted on the dataset of CSDMC2010 Spam corpus. The results for the considered dataset are evaluated using individual classifiers, bagging, and boosting ensemble approaches. The system performance is accessed in terms of precision, recall, f-measure, and accuracy. The experimental outcomes indicates the distinguish results for the detection of email spam using ensemble methods.

***Keywords*: Email Spam Detection, Ensemble Methods, Bagging, Boosting, Multinomial Naïve Bayes, and J48 Decision Tree.**

I. INTRODUCTION

The advancements in the field of technology, internet, and smart phones have increased the accessibility of electronic gadgets for the human beings. These advancements have made the long distance social communication handy. The electronic mail (email) system is one among the widely used interaction source for the official and business perspectives. The emails are lesser used for the personal usage. Although the email system is convenient and cheap source, the increasing spam activities have made it expensive. Spam emails are the unwanted undesired emails that can harm the legitimate users. Email spam can be in the form of business advertisement, promotional activities, or any other kind of redirection URLs that can lead to any financial laundering, malicious, virus, and other kinds of threatening activities. The filtration of email spam process leads to wastage of several hours and energy of human. On the daily basis, an average official person receives 121 emails per day [1]. This number is increasing regularly due to enhancement in the small scale entrepreneur companies and their advertisements. This rapid growth of users is expected to grow upto 4.3 billion till the year 2023 [1]. The increase in the number of users also arise the probability of email spam. The spam can affect the legitimate users adversely by adapting the identity of legitimate users and can deliver the shortened phishing URLs. The concept of email spam is not the new; it began in the year 1978 by Thuerk by sending the advertisement based spam to 400 users. The latest statistics of email spam indicate the increasing ratio of spam email with respective to total number of emails received. The data available for the August, 2019 indicates the average daily legitimate and spam emails are 72.16 billion and 416.04 billion respectively. The percentage of this statistics is shown in fig. 1. This statistics indicate the availability of 85.22% email spam is available in the total received emails. The percentage of email spam is continuously increasing as indicated from the statistics available for last one year (September, 2018 to August 2019) [2]. This statistics is available is presented in table I. The statistics discussed in table I indicates the receiving of more than 80% of emails are spam among the total number of emails received.

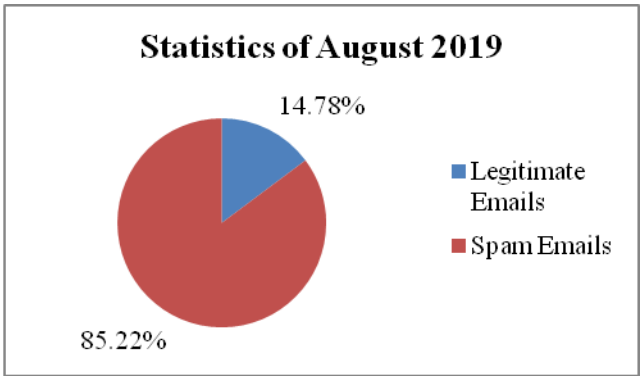

Fig. 1: Latest email statistics of August, 2019

Table I: Email statistics from Sep 2018 to Aug 2019

Month, Year	Total number of emails (in Billion)	Total number of spam emails (in Billion)
Sep 2018	354.5	301.95
Nov 2018	302.2	257.75
Feb 2019	239.22	204.19
Apr 2019	489.34	416.78
Jun 2019	539.22	459.4
Aug 2019	488.2	416.04

Manuscript published on 30 September 2019
* Correspondence Author
*** Uma Bhardwaj**, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India.
Email: umabhardwaj90@gmail.com
Priti Sharma, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India.
Email: pritish80@yahoo.co.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

These increasing spamming activities in the field of emails demands to consider the filtration and detection approaches that can easily detect the email spam.

Despite there is the availability of a lot of machine learning methods to detect the email spam, the individual methods lacks somewhere. In this research work, ensemble methods of bagging and boosting are considered to detect the email spam. The machine learning methods of Multinomial Naïve Bayes (MNB) and J48 Decision Tree are considered and ensembled. The experimentation is conducted on the dataset of CSDMC2010 Spam corpus for the methods of individual classifiers (MNB and J48), bagging and boosting (AdaBoost). The bagging method works by slitting the data into sets and procedure is conducted in parallel sequence. This reduces variance and saves time. Boosting approach has more advantageous features as it combines the properties of both the classifiers and avoids their lacking features. Moreover, it also reduces the variance along with the reduction in the biasness. The boosting approach also prioritizes the unclassified dataset samples by adding the additional weight to them. The performance of the proposed ensemble approaches is accessed in terms of evaluation metrics results.

The other sections of the manuscript are arranged as illustrated: Section 2 briefs the existing research studies relevant to the email spam detection. Section 3 presents the proposed research methodology with classification methods of individual classifiers, bagging, and boosting concept for email spam detection. Section 4 presents the result outcomes and discussion on evaluated results. The comparison of all the individual classifiers, bagging, and boosting approaches is also discussed in this section. And the paper ends with the conclusion as illustrated in Section 5.

II. Related Work

Email spam acts as the plague of networking technology as it affects the individual as well as business users with the receiving of multiple spam emails. The email spam are available in various formats such as shortened URLs, attachments, advertisements, spyware, Trojans, etc. The researchers are continuously working on the techniques to control and reduce the email spam. In this research paper, we have discussed the latest and quality publications of researchers who have presented some impactful work for the email spam detection. The considered research publications are discussed here.

Hassan [3] has tested the combinational approach of k-means clustering with the supervised learning classifiers. The author has performed the research experimentation to check the change in the evaluation accuracy results with the addition of clustering approach with classifiers. The research experimentation was conducted on the Enron-Spam dataset. The combined approached includes the clustering with supervised classifiers of decision tree (DT), logistic regression (LR), k-nearest-neighbor (KNN), support vector machine (SVM), and naïve bayes (NB). The evaluation results indicated that there was no much change in the accuracy was observed. Only a little improvement for the combined clustering and LR approach is noted. All the methods were unaffected.

Olatunji [4] has performed the detection of email spam by considering Extreme Learning Machines (ELM) and SVM approach. In the research experimentation, ELM approach consumes lesser time as compared to SVM algorithm but SVM outperformed in terms of accuracy as compared to ELM. The concepts of ELM & SVM are also compared with Fuzzy logic, BART, NSA, PSO & NSA-PSO concept and shows better results than others for the Email spam classification.

Kumaresan and Palanisamy [5] adapted the algorithm of Cuckoo Search by adding the StepSize feature to it and introduced as Cuckoo Search with Stepsize (SCS) approach. In this work, authors have also used the classification approach of SVM. Here, feature selection is performed with proposed SCS approach and classification with SVM approach. The performance efficacy results indicate the dominance of proposed SCS-SVM approach as compared to existing CS-SVM.

Chawathe [6] has used the fuzzy rule (FURIA) to improve the security of email system. Here, fuzzy rules based system is designed to detect the email spam and database of SpamBase is used for the experimentation. The author has conducted the experimentation on java virtual machine and Weka tool. There was the also the consideration of PART and JRip algorithms for the classification comparison. The author observed the comparable performance of the proposed concepts with other concepts.

Cohen et al. [7] have used extensive email features related to email header, content, and other attachments to analyze the malicious email content. There is the manual collection of email dataset by analyzing the emails on VirusTotal tool to define the status of emails. Different experiments based on feature set, depth of emails with the analysis of spamming affects, etc. were conducted by authors. The authors have used several machine learning classifiers to analyze the results with integrated detection rate parameter and reported effective evaluation using random forest classifier.

Agarwal and Kumar [8] have amalgamated the concepts of Particle Swarm Optimization (PSO) and Naïve Bayes algorithm to detect the spam emails. In this approach, the PSO algorithm used to improve the classification efficacy evaluated using Naïve Bayes approach. The authors have considered the ling spam corpus and performed the experimentation by randomly selected 1000 emails. The evaluation of system was conducted in terms of evaluation metrics. The integrated approach indicates the superiority of concept as compared to individual Naïve Bayes classifier.

Naem et al. [9] have conducted the experimentation of email spam detection by combining the concepts of boosting approach and Ant-Lion Optimization (ALO). The combined approach was termed as ALO-Boosting and experimented on the datasets of SpamAssassin and CSDMC2010. In the combined approach, the ALO algorithm is utilized for the selection of features and Boosting approach is utilized for the classification procedure. The proposed integrated approach was compared with the classifiers of ALO-Bagging, ALO-SVM, ALO-KNN, Boosting, Bagging, and individual methods SVM & KNN.

The proposed approach indicate the efficacy of mentioned proposed approach as compared to other concepts in terms of evaluation measures.

Gupta et al. [10] have considered the ensemble learning approach for the classification of email and SMS spam using the machine learning classifiers of the Decision Tree (DT), Bernoulli Naïve Bayes (BNB), Multinomial Naïve Bayes (MNB), and Gaussian Naïve Bayes (GNB). The authors have used the voting ensemble for the classification with different combinations of mentioned classifiers. The experimentation was conducted on the dataset collected from the UCI website. The authors have achieved higher performance accuracy result with a combination of DT, BNB, & GNB classifiers in case of SMS dataset and a combination of all the classifiers in case of email dataset.

III. Research Methodology

The research methodology involves the steps of consideration of research dataset, pre-processing of dataset, feature extraction & selection, and classification. The workflow of the research methodology is presented in fig. 2 and the step by step algorithm is discussed further.

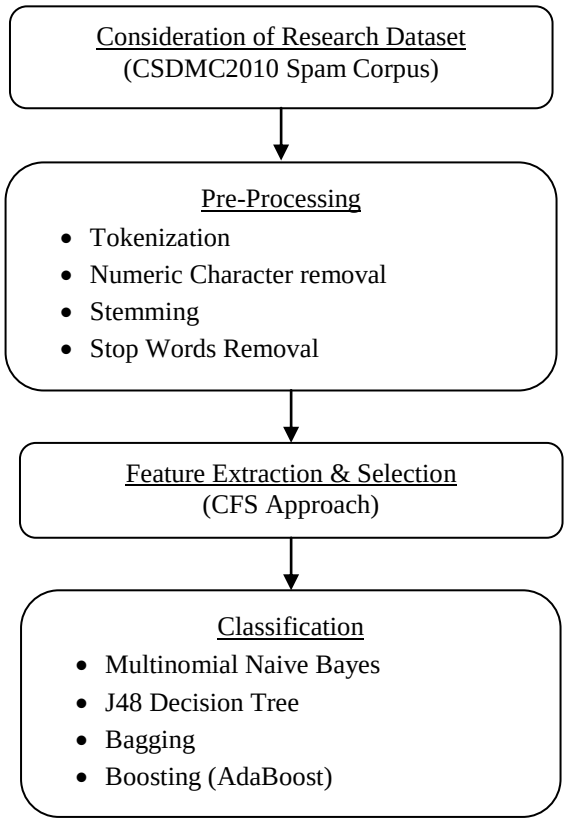

Fig. 2: The Proposed Workflow of Email Spam Detection

A. Consideration of Research Dataset

The collection and consideration of dataset is the first step of email spam detection. Here, the research experimentation is conducted on the dataset of CSDMC2010 Spam Corpus. CSDMC2010 dataset is a dataset of emails which contains 1378 spam emails and 2949 legitimate emails. This dataset was derived and presented during the completion of data mining datasets in International Conference on Neural Information Processing (ICONIP 2010) organized in Sydney, Australia. The dataset contains a total number of 4327 emails. The next step of emails spam detection approach is pre-processing.

B. Pre-processing

The considered dataset needs to be processed before further experimentation. The pre-processing involves the steps of tokenization, removal of numeric digits, stemming, and stop word removal. The first phase of pre-processing is tokenization. Tokenization process split the words based on the space achieved after each word. The tokenized words are further referred as tokens. These tokens are further checked for the word type as numeric or string characters. The numeric digits are removed from the dataset to reduce the search space as there is no usage of these digits. The remaining processed data further considered for the phase of stemming. Stemming is the process to convert the tokens into the root word in case there is the attachment of any kind of prefix or suffix with the root word. For the instance the possible words for the word 'adjust' are 'adjustment', 'adjustable', 'adjusts', 'adjusted ', 'adjusting', 'adjustor', 'adjustability', etc. The process of stemming removes the prefix and suffix from these words such as 'ment', 'able', 'ed', 's', 'ing', 'or', and 'ability' respectively. The process of stemming not only reduces the search space by reducing the matching depth but also enhances the efficiency of system to detect the spamming keywords. This pre-processed data is further considered for the final phase of stop words removal process. The stop words are the helping words of English language that complete the grammatical syntax. These words are unused to detect the email spam and can be removed. The instances of stop words involves the words such as 'the', 'a', 'are', 'their', 'per', 'on', 'until', 'why', etc. The stop words removal procedure reduces the search space upto a great extent. This pre-processed data is used for the step of feature extraction and selection.

C. Feature Extraction and Selection

In the overall procedure of email spam detection, there is the major role played by features available in the dataset. The features extracted from the dataset are relevant to the length of the document, word frequency, numeric characters, frequency of spam words, frequency of legitimate words, probability of inappropriate words, etc. The feature components are checked for the token frequency using the term frequency method. The tokens above the threshold are considered into account and lower frequency tokens are discarded. This process also reduces the search space. Further, the extracted features are further selected using the correlation based feature selection (CFS) method. CFS method makes the selection based on the most relevant features for any specific class. If there are number of classes & features *c* & *k* respectively with feature set *f*, then CFS method can be formulated mathematically as illustrated in (1).

$$CFS = max_{S_k}\left[\frac{r_{cf_1}, r_{cf_2}, r_{cf_3}, \ldots, r_{cf_k}}{\sqrt{k+2\left(r_{f_1f_2}+\ldots r_{f_if_j}+\ldots r_{f_kf_1}\right)}}\right] \quad (1)$$

Here r_{cf} is the mean correlation feature class value, r_{ff} is the mean correlation feature-feature value.

The selected features using the method of CFS are further considered for the classification and final detection of email spam.

D. Classification

The classification of dataset emails into their respective class of spam and legitimate is the final step of email spam detection. In this research, the classification is conducted initially using the individual classifiers of Multinomial Naïve Bayes classifiers and then ensemble methods of bagging and boosting are applied. All the above methods for classification are discussed below.

D(1). Multinomial Naïve Bayes

Multinomial naïve bayes classifier is based on the bayes theorem [11]. The algorithm possesses multinomial distribution with respect to number of occurrences of word in text document. The Multinomial Naïve Bayes classifier considers the assumption of strong independence and works efficiently as supervised learning approach. In this classifier, the word vector count represents the data. The concept of multinomial naïve bayes classifier is further elaborated by considering an example of an email which contains the word 'winner'. The probability of word 'winner' as a spam class can be evaluated using (2).

$$p(s|w) = \frac{p(w|s)\,p(s)}{p(w|s)\,p(s) + p(w|l)p(l)} \quad (2)$$

In (1), the word 'winner' is represented by w, the email classes of legitimate and spam are expressed with symbols l and s respectively. The possibility of an email belongs to spam class with 'winner' word is represented by $p(s|w)$ whose value dependent to the $p(s)$, $p(w|s)$, $p(l)$ and $p(w|l)$.

D(2). J48 Decision Tree

The decision tree (DT) algorithm is tree based hierarchical structure that possesses terminal nodes as decision outcomes and non-terminal nodes as test attributes. The popular DT algorithms are C4.5, ID3, CHAID, and classification & regression tree. Among these versions, decision tree J48 (C4.5) is the refined version of ID3 and possesses the efficient classification accuracy.

The J48 decision tree [12] splits the data in the form of various different sets with the ability so that a decision can be made with each feature attribute. J48 Decision tree uses the entropy function to generate rules of decision tree with the help of target emails database. There is the recursive working of algorithm until the processing and categorization of each data attribute. The classification of emails can be evaluated as described in (3).

$$Entropy\ (Email) = -\sum_{j=1}^{n} \frac{|Email_j|}{|Email|} log_2 \frac{|Email_j|}{|Email|} \quad (3)$$

Where, *Email* is the n-gram function which can be unigram, bigram, and trigram as per the considered cases. Entropy evaluates the prediction of email as the spam or legitimate email with the concept of J48 decision tree algorithm.

D(3). Bagging Approach

The bagging approach is an ensemble method considered for the classification. The bagging approach works in the parallel manner by splitting the overall dataset into the sets and assign to classifier. Here, Multinomial Naïve Bayes (MNB) and J48 Decision Tree are considered for the bagging approach. The bagging approach reduces the variance for the classification errors and also takes lesser time for the computation of email spam as it work in the parallel manner. The overall data samples are splitted among the considered classifiers of MNB and J48. The system result is the average of the result evaluated using the classification algorithms.

D(4). Boosting Approach

The boosting approach is also an ensemble method that can overcome the drawback of a classifier by adding the properties from another classifier. The boosting involves the series based process to initially classify the training samples using first model, further introducing the second model to rectify the errors remains in the first model and continues until the perfect rectification of training samples. Boosting reduces both the bias and variance of the classification and improves the classification results. The boosting approach is sensitive to noise and the computation time of boosting process is more as compared to another ensemble approach of bagging. In this research work, Adaboost is used for the boosting of naïve bayes and J48 decision tree classifiers. Adaboost was the first successful method developed to boost the binary classification. The Adaboost adapts the boosting property to focus more on misclassified data samples by adding more weights to the samples which are found to be misclassified. The procedure proceeds based on iterations in which each sample is checked after each iteration and misclassified samples are prioritized by adding more weight to further classify that sample appropriately. The overall classification is the weighted sum of all the ensemble predictions.

IV. Results and Discussions

The results of the proposed ensemble methods are accessed in terms of evaluation measures of accuracy, recall, precision, and f-measure. The experimentation is performed on the Java based Eclipse IDE simulation software. The system configuration is Window based operating system with 8GB of RAM, and Intel i5-core processor.

Table II: Confusion Matrix for Email Spam Detection

		Ground Truth	
		Spam Emails	**Legitimate Emails**
Predicted Results	**Spam Emails**	TP = a	FP = d
	Legitimate Emails	FN = c	TN = b

The evaluation measures of accuracy, recall, precision, and f-measure are calculated using

the confusion matrix along the values of False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP).

In this research work, TP is considered as the number of emails under evaluation predicted as spam using ensemble method is same as per the actual value of email. TN is considered as the number of emails evaluated as legitimate using ensemble method is same as per the actual value of those emails. FP is considered as the number of emails evaluated as spam using ensemble method and the actual value of email is legitimate. FN indicates the number of emails evaluated as legitimate using ensemble method but the actual value of those emails is spam. The considered confusion matrix is illustrated in table II.

Based on the values of FN, FP, TN, and TP, the formulations of accuracy, recall, precision, and f-measure are illustrated in (4)-(7).

$$Accuracy = \frac{a+b}{a+b+c+d} \quad (4)$$

$$Recall = \frac{a}{a+c} \quad (5)$$

$$Precision = \frac{a}{a+d} \quad (6)$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

To evaluate the results using the ensemble methods, it is necessary to train the system before testing. The training of the system is performed based on the email samples available in the dataset. Among the available data samples, 80% of the data is used for the training and 20% data used for the testing. The results are determined based on the testing data emails. As per the 20% of the dataset, there are 276 spam emails and 590 are the legitimate emails on which performance of the system is determined. The evaluated values of FN, FP, TN, and TP are illustrated in table III.

Table III: Evaluated Parameters of FN, FP, TN, and TP

	MNB	J48	Bagging	Boosting
TP	236	249	267	272
FN	40	27	09	04
FP	41	32	13	06
TN	549	558	577	584

From table III, it can seen that ensemble methods of bagging and boosting approaches have lesser FP and FN values as compared to individual classifiers which indicate the better result values of ensemble methods. Based on the evaluated values of FN, FP, TN, and TP (refer to table III), the evaluation metrics of accuracy, recall, precision, and f-measure are determined as illustrated in table IV.

Table IV: Performance Evaluation Results

	MNB	J48	Bagging	Boosting
Precision (%)	85.20	88.61	95.36	97.84
Recall (%)	85.51	90.22	96.74	98.55
F-Measure (%)	85.35	89.41	96.05	98.19
Accuracy (%)	90.65	93.19	97.46	98.85

In case of individual classifiers of MNB and J48, the methods have achieved the classification accuracy of 90.6% and 93.19% respectively. This indicates the better result of the J48 algorithm as compared to MNB approach. In another case of ensemble methods, bagging and boosting approach have accomplished the accuracy values of 97.46% and 98.85% which points to the better efficacy of boosting approach as compared to bagging.

These evaluated results illustrated in table IV are further considered for the graphical illustration in fig. 3 for the evaluation measures of precision, recall, f-measure, and accuracy parameter.

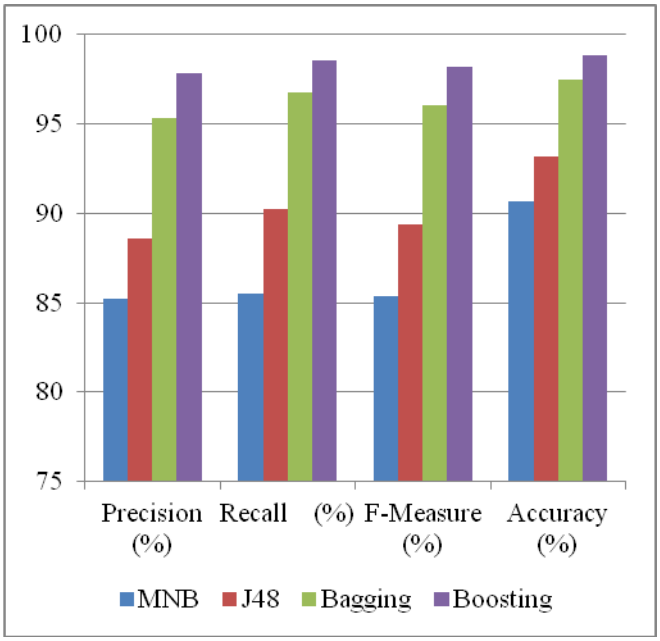

Fig. 3: Comparison based on Evaluation Parameters

The result values illustrated in fig. 3 clearly indicate the better performance of boosting approach in all the parametric measures of precision, recall, f-measure, and accuracy. The MNB approach has achieved the lower performance result. Then, J48 algorithm have achieved the better than MNB but lacks than the results accomplished with ensemble methods. In the overall scenario, ensemble methods performed better than individual classifiers.

V. CONCLUSION

The email spam is the one among the challenging issues in the field of computing technologies. The methods of machine learning are significant to detect the spamming activities. Despite the significance of machine leaning methods to detect the email spam, there are some drawbacks of the methods are observed. To overcome the lacking features of methods to detect the email spam, this research work have brought the ensemble methods for the improvement in the results of individual classifiers to detect the email spam. The

methods of Multinomial Naïve Bayes and J48 Decision Tree algorithms are ensembled to tackle the problem. The first ensemble method considered was bagging approach that works in the parallel manner by splitting the data samples to considered classifiers.

Another approach was boosting approach that works in the series manner with the combined properties of both the classifiers. The research experimentation has been conducted based on the dataset of CSDMC2010 Spam Corpus. The evaluated experimental results in terms of precision, recall, f-measure, and accuracy indicate the superior performance of ensemble methods as compared to individual classifiers of J48 and MNB. In the comprehensive manner, boosting approach dominates with accuracy value of 98.85%, then bagging approach achieved accuracy of 97.46%, further J48 approach attained accuracy of 93.19% and MNB value lacks with accuracy of 90.65% respectively.

REFERENCES

1. H. Tschabitscher, 2019, “Fascinating Email Facts and Statistics”, Online Available at: https://www.lifewire.com/how-many-emails-are-sent-every-day-1171210.
2. Talos Intelligence Report, 2019. “Total Global Email & Spam Volume For August 2019”, available online at: https://www.talosintelligence.com/reputation_center/email_rep, Last accessed 03 Sep 2019.
3. D. Hassan, "Investigating the effect of combining text clustering with classification on improving spam email detection." In: *Proc. of International Conference on Intelligent Systems Design and Applications*, pp. 99-107. Springer, Cham, 2016.
4. S. O. Olatunji, "Extreme Learning machines and Support Vector Machines models for email spam detection", In: *Proc. of IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Windsor, pp.1-6, 2017.
5. T. Kumaresan and C. Palanisamy, "Email spam classification using S-cuckoo search and support vector machine", *International Journal of Bio-Inspired Computation*, Vol.9, No.3, pp.142-156, 2017.
6. S. Chawathe, "Improving Email Security with Fuzzy Rules", In: *Proc. of 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, New York, pp.1864-1869, 2018.
7. A. Cohen, N. Nissim, and Y. Elovici, "Novel Set of General Descriptive Features For Enhanced Detection of Malicious Emails Using Machine Learning Methods", *Expert Systems with Applications*, Vol.110, pp.1463-169, 2018.
8. K. Agarwal, and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization." In: *Proc. of the Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 685-690. IEEE, 2018.
9. A. A. Naem, N. I. Ghali, and A. A. Saleh, "Antlion optimization and boosting classifier for spam email detection." *Future Computing and Informatics Journal*, Vol. 3, No. 2, pp. 436-442, 2018.
10. V. Gupta, A. Mehta, A. Goel, U. Dixit, and A. C. Pandey, "Spam Detection Using Ensemble Learning." In *Harmony Search and Nature Inspired Optimization Algorithms*, pp. 661-668. Springer, Singapore, 2019.
11. A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited", In: *Proc. of the Australasian Joint Conference on Artificial Intelligence*, Vol.3339, pp. 488-499, Cairns, Australia, 2004.
12. S. Youn, and D. McLeod, "A comparative study for email classification", In: *Proc. of the Advances and innovations in systems, computing sciences and software engineering*, pp.387-391, Springer, Dordrecht, 2007.

AUTHORS PROFILE

Uma Bhardwaj received B.C.A. in 2010 and M.C.A. in 2013 from Maharshi Dayanand University, Rohtak. She is a research scholar in Maharshi Dayanand University Rohtak in Department of Computer Science & Applications. Her area of research is "spam - ham mail classification ". Her research interest includes Data Mining, Text Mining, Character Recognition and Natural language processing.

Priti Sharma is assistant professor in Department of Computer Science & Applications in Maharshi Dayanand University, Rohtak. She received Ph.D. in Computer Science from Kurukshetra University, Kurukshetra. Her research interest includes Software Engineering, Software Re-engineering, Data Mining, and Software Metrics. She has teaching experience of 10 years having approx. 40 publications.