International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Detection of Email Spam using an Ensemble based Boosting Technique

Uma Bhardwaj, Priti Sharma

Abstract*: Email is amongst the advanced socializing communication source that is mostly adapted by business and commercial users. Although there are various assets of email as it is quick, cheap, and efficient communication resource, the service is misused by various users for the personal or professional purposes by spreading the useless and extra emails which are termed as email spam. There is the existing research of authors who have used machine learning methods for the detection of email spam and achieved effective precision values, but the individual methods mostly noted with some shortcomings which leads to lack in the performance. In this research work, ensemble based boosting technique on machine learning classifiers of Multinomial Naïve Bayes and J48 classifiers is proposed for the detection of email spam. The system detects email spam by adding the strong features of one classifier into another with the help of Adaboost algorithm. The results of the proposed ensemble technique are evaluated with accuracy and sensitivity parameters. The evaluated results indicate the effectiveness of proposed concept to detect the email spam in comparison with individual methods and other considered existing concepts.

***Index Terms*: Adaboost, Boosting, Decision tree J48 algorithm, Email spam, Multinomial Naïve Bayes, Spam filtration, Text analysis.**

I. INTRODUCTION

In this internet era, the availability of social media networks and applications has made the communication among humans very advanced. These advancements in the field of communication and technology have converted the socializing means from traditional short messages, radio messages, telegraphs etc. to advanced methods such as email, video streaming, and social networking applications. After the invention of email based communication system in 1971, emails were adapted as the preferable communication source especially by official and business users. Due to easy accessibility and cost effectiveness of email system, it was started misusing by sending number of useless and extra emails which are termed as email spam. The first email spam was performed by Gary Thuerk in 1978 to seek the attention of people about the introduction of their DECSYSTEM computer products. This email spam was sent to 400 people among the 2600 people on ARPAnet [1].

Email spam acts as the plague of networking technology as it affects the each individual user to multinational organizations with the receiving of multiple spam emails. The email spam not only includes the bulk unwanted useless emails but it is also a type of spreading the various spyware, Trojans, worms and viruses etc. These can be sent as an attachment or a redirection link. Another category of email spam is the blocking network with huge traffic and denial of service attacks. These lead to reducing the internet speed and data interruption. Due to all these spam, both the email service providers and internet service providers are adversely impacted. There are also the instances of email spam where disabled users are affected. Visually and Hearing impaired users use different aid devices to understand the emails. But after receiving the lot of spam emails, it becomes difficult for them to handle the useful emails.

Another category of affected users are the working employees whose lot of time and energy got wasted to filter out the legitimate and useful emails out of spamming emails. The professionals waste a lot of money and time to get out of email spam. Spammers spread the email spam without revealing their identity with the use of different user address. The total number of users and emails are also increasing with the increase of awareness in people about the use of email system. As per the Radicati Group report [2], the numbers of worldwide email users were 3.8 billion in 2018 which is expected to expand the number upto 4.2 billion till the end of year 2022. The numbers of worldwide sent and received emails were also 281 billion in 2018 which is expected to expand the number upto 333 billion in 2022 [2]. As per the Talos Intelligence report of December, 2018 [3], the average daily number of email data was reported 364.25 billion which contain 311.24 billion spam emails. The statistics of December 2018 for the email spam is presented in fig. 1. Moreover, the statistics of spam emails, legitimate emails, and total number of emails from July 2018 to December 2018 is presented in fig. 2. The statistics illustrated in fig. 1 and fig. 2 indicates that there is higher availability of spam emails in comparison with legitimate emails. In each case, the count of spam emails is above 75%.

There are various existing methods based on machine learning approaches and other concepts for the detection of spam content (as illustrated in related work section). The individual methods are noted with some shortcomings which lead to lack in the performance. In this research work, machine learning based Multinomial Naïve Bayes and Decision Tree J48 classifiers are ensembled and boosting approach is applied with the help of Adaboost to detect the email spam. In this boosting approach, the weakness of one classifiers is replaced with another and vice-versa to achieve the overall effective performance results.

Manuscript published on 30 September 2019.
***Correspondence Author(s)**
Uma Bhardwaj, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India.
Priti Sharma, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India.

The experimentation of the ensembled approach is performed on Ling Spam Corpus. The other sections of the manuscript are arranges as illustrated: Section 2 briefs the existing research studies relevant to the email spam detection. Section 3 discusses the preliminaries of Multinomial Naïve Bayes, Decision Tree J48 classifier, and Boosting concept. Section 4 presents the research methodology of boosting concept for email spam detection. Section 5 presents the considered database, results and discussion. The comparison of used boosting approach with other concepts is also discussed in this section 5. And the paper ends with the conclusion as illustrated in Section 6.

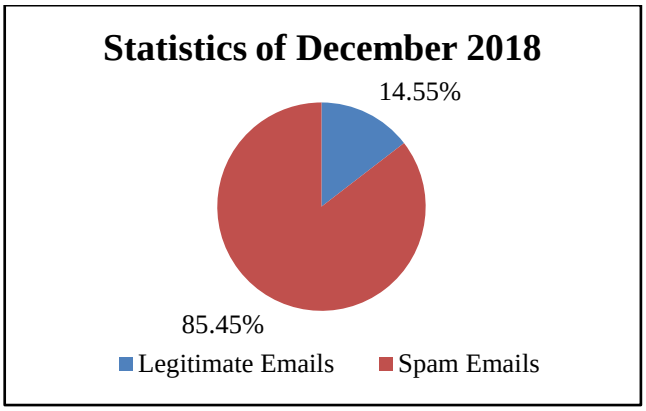

Fig. 1: Email Statistics of December 2018 [3]

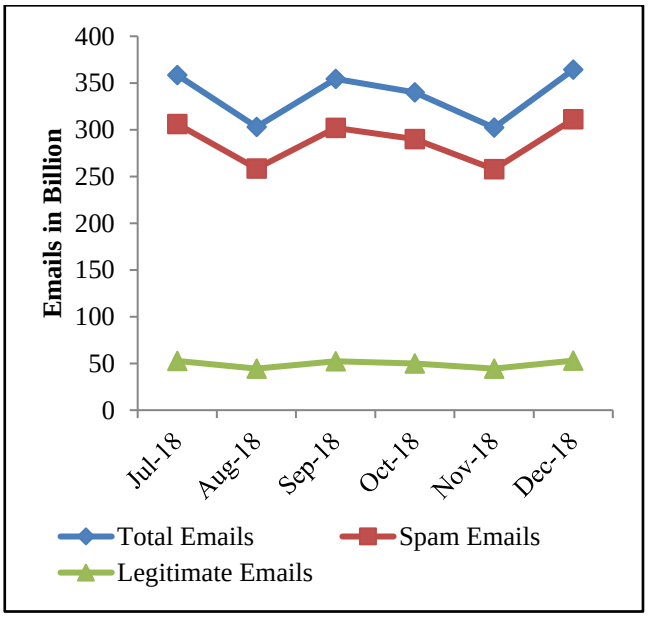

Fig. 2: Email Statistics from July 2018 to December 2018 [3]

II. RELATED WORK

The growth in the email communication facilities has also enhances the threatening users who are responsible for spreading email spam. There are various existing methods and concepts that are exploited by researchers to detect the email spam. The selected latest and quality publication work related to email spam detection is discussed in this section.

Renuka and Visalakshi [4] have performed the classification and detection of email spam by considering Support vector Machine (SVM). The authors have also used the feature selection approach of Latent Semantic Indexing (LSI) to enhance the system performance. Dataset of Ling Spam Email Corpus is used.

Tuteja and Bogiri [5] have created the manual dataset for application of email spam detection. The authors have adapted the classification approach of ANN. In this process, feature extraction is performed with K-means clustering, dataset training is performed with Back Propagation Neural Network, and final classification & email spam detection is performed with Feed Forward Neural Network. Kaur and Sharma [6] have amalgamated the methods of Decision Tree approach with PSO algorithm for the improvements. The results of proposed integrated concept are compared to k-means & SVM approach with & without unsupervised filtration on the basis of evaluation parameters. The work lacks as the authors have not mentioned about any feature extraction approach in their research manuscript.

Kumaresan and Palanisamy [7] 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. Further, Olatunji et al. [8] have 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. Further, Chawathe [9] have 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. Author observed the comparable performance of the proposed concepts with other concepts. Cohen et al. [10] have used extensive email features related to email header, content, and other attachments to analyze the malicious email content. Authors have used several machine learning classifiers to analyze the results with integrated detection rate parameter and reported effective evaluation using random forest classifier. Furthermore, Diale et al. [11] have focused on the feature set evaluation with cosine similarity and auto-encoder by considering distributed memory and distributed bag of words to generate the vector space model for feature representation and evaluation. The classification on Enron corpus is performed using SVM, C4.5 decision tree, and random forest.

III. PRELIMINARIES

This research work considers the machine learning based Multinomial Naive Bayes and Decision Tree J48 classifiers for boosting with Adaboost approach. The basic of these concepts is discussed here.

Retrieval Number: K13650981119/19©BEIESP
DOI: 10.35940/ijitee.K1365.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

A. Multinomial Naïve Bayes

Multinomial naïve bayes classifier is based on the bayes theorem. 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 (1).

$$p(s|w) = \frac{p(W|S)\,p(s)}{p(W|S)\,p(s) + p(W|\bar{l})p(l)} \quad (1)$$

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 $p(w|l)$.

B. Decision Tree J48

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 decision tree J48 splits the data in the form of various different sets with the ability so that a decision can be made with each feature attribute. Decision tree J48 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 (2).

$$Entropy\ (Email) = -\sum_{j=1}^{n} \frac{|Email_j|}{|Email|} log_2 \frac{|Email_j|}{|Email|} \quad (2)$$

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 decision tree J48 algorithm.

C. Boosting Approach

The boosting approach is the 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. RESEARCH METHODOLOGY

The research methodology of email spam detection involves three major steps of pre-processing of database, feature extraction and selection, and classification. The workflow of research methodology is discussed in fig. 3. The thorough explanation of algorithm is also explained as follows.

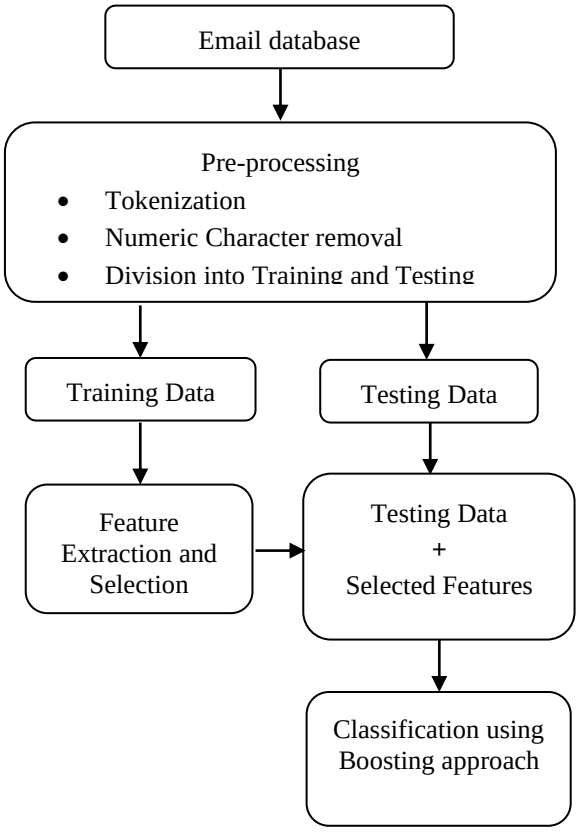

Fig. 3: The Working Procedure of Proposed Ensemble based Boosting Technique

A. Pre-Processing

The research methodology of proposed system begins with the pre-processing of email database. It composed of tokenization, numeric character removal, and categorization of email database as training and testing. In this research work, we are working on the textual database based email spam detection. This makes to remove the email database containing imagery related spam. Then, the tokenization of the corpus is performed in which each word of email corpus is considered as the individual token. Each email is represented in the following manner with (3).

$$email_i = (t_{i1}, t_{i2}, \dots, t_{im}) \quad (3)$$

Where, t_{ij} describe the token frequency for the token t_j in $email_i$. These values are evaluated in terms of binary frequency order.

These pre-processed tokenized emails are considered to remove the numeric digits from the textual data to decrease the search space. Further, the email database is categorized for the training and testing. For training and testing purpose, overall database is categorized into the ratio of 80:20, where 80% database is exploited for training and 20% for testing purpose. The next step is feature extraction and selection.

B. Feature Extraction and Selection

There is the major role of considered feature set to detect the email spam. There are total 258 features are extracted for the classification. These features are related to length of the document, word frequency, numeric characters, frequency of spam words, frequency of legitimate words, probability of inappropriate words, etc. The initial step is to consider the words of the data as tokens. All the available token frequency per document is determined by using term frequency method. The considered word frequency is checked for the threshold and lesser frequency words are eradicated. The eradication of useless words reduces the search space and improves the importance of other words. There is further reduction of features using correlation based feature selection (CFS) method to keep only the most useful features. 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 (4).

$$CFS = max_{S_k}\left[\frac{r_{cf_1}, r_{cf_2}, r_{cf_3}, \dots, r_{cf_k}}{\sqrt{k+2\left(r_{f_1f_2} + \dots r_{f_if_j} + \dots r_{f_kf_1}\right)}}\right] \quad (4)$$

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

These extracted and selected features are used to train the model and classification of testing data is performed based on the trained model.

C. Classification

The classification of email spam database is performed using Adaboost boosting approach. In this research work, Adaboost is used for the boosting of classifiers multinomial naïve bayes and decision tree J48 classifiers. In the individual case of Multinomial Naïve Bayes algorithm, it lacks for the classification of contextual emails. Adaboost recovered this drawback with the use of decision tree J48 algorithm. Moreover, individual decision tree J48 also lacks in case of noisy and long email. This increases the space complexity and makes the classification process slower. To overcome this drawback, Adaboost adds the property of multinomial naïve bayes approach. Adaboost works on the assumption to focus more on misclassified instances by increasing its weight value.

V. EXPERIMENTAL RESULTS

The experimental results of the Adaboost boosting algorithm are evaluated for the research database of Ling Spam Corpus. The experimentation is conducted with window operating system with configuration of 4 GB RAM, and Intel i5 processor. The results are simulated on the java based Eclipse IDE simulation software. The evaluation metrics, considered database and evaluated results along with the comparison are illustrated in this section.

A. Ling Spam Corpus

The concept of Adaboost boosting approach is experimented using Ling Spam Corpus. The research database consists of legitimate and spam emails collected from the different scientific and professional linguistics by Androutsopoulos et al. in the year 2000 [12]. In this research work, we have performed the experimentation on bare and ling category of ling spam corpus. In bare category of database, both the stop-word list and lemmatiser are disabled. In lemm category of database, stop word list is disabled but lemmatiser is enabled. The presence of lemmatiser helps to reduce the search space of database. Both categories consist of 10 folders which contains legitimate and spam emails. Both categories contain total 2412 legitimate emails and 481 spam emails which make a total of 2893 emails. Spam emails are available with the name containing "spmsg" and all other are legitimate (ham) emails.

B. Evaluation Metrics

The results of the boosting approach are evaluated in terms of accuracy and sensitivity. The formulation of these parameters depends on the value 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 boosting approach is same as per the actual value of email. TN is considered as the number of emails evaluated as legitimate is same as per the actual value of those emails. FP is considered as the number of emails evaluated as spam using boosting approach and the actual value of email is legitimate. FN indicates the number of emails evaluated as legitimate using boosting approach but the actual value of those emails is spam. The formulation of accuracy and sensitivity are discussed here.

Sensitivity indicates the possibility of actual recognition of email spam. It is the positive labeled data returned by the system classifier out of total class data. Sensitivity is also popular with the names True Positive Rate and Recall. The formulation of the sensitivity is illustrated in (5).

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

Accuracy indicates the ratio of positive predictions to count of overall data values. The formulation of the accuracy is illustrated in (6).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

C. Results and Comparison

The results of the boosting approach are evaluated for the both categories of ling spam corpus in terms of considered evaluation metrics.

Retrieval Number: K13650981119/19©BEIESP
DOI: 10.35940/ijitee.K1365.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Among the total ten folders of emails in each category, eight folders used for training and two folders of emails are considered for testing. In these two folders, there are total 580 emails which contain 483 legitimate emails and 97 spam emails. The testing results are evaluated based on these email statistics.

C (1). Bare category

The bare category consists of both the stop word list and lemmatiser disabled. The evaluated results of bare category using boosting approach are illustrated in table I along with the results evaluated using individual multinomial naive bayes and decision tree J48 approach.

Table I: Evaluated Results for Bare Category

	Multinomial Naïve Bayes	Decision Tree J48	Proposed boosting technique
TP	61	67	77
FN	36	30	20
FP	68	56	26
TN	415	427	457
Sensitivity	62.89 %	69.07 %	79.38 %
Accuracy	82.07 %	85.17 %	92.07 %

C (2). Lemm category

Lemm category consists of enabled lemmatiser but stop word list disabled. The calculated results of lemm category using boosting approach, individual multinomial naive bayes, and decision tree J48 are illustrated in table II.

Table II: Evaluated Results for Lemm Category

	Multinomial Naïve Bayes	Decision Tree J48	Proposed boosting technique
TP	68	70	86
FN	29	27	11
FP	54	42	05
TN	429	441	478
Sensitivity	70.10 %	72.16 %	88.66 %
Accuracy	85.69 %	88.10 %	97.24 %

The evaluated results illustrated in table I and table II in terms of sensitivity and accuracy indicates that the boosting approach outperformed than the individual results of multinomial naive bayes and decision tree J48 approach. The evaluated results of bare and lemm category using boosting approach are further compared with existing algorithm of Stepsize Cuckoo Search (SCS) [7] for both the bare and lemm category. The comparison results for bare and lemm categories are illustrated by fig. 4 and fig. 5 respectively.

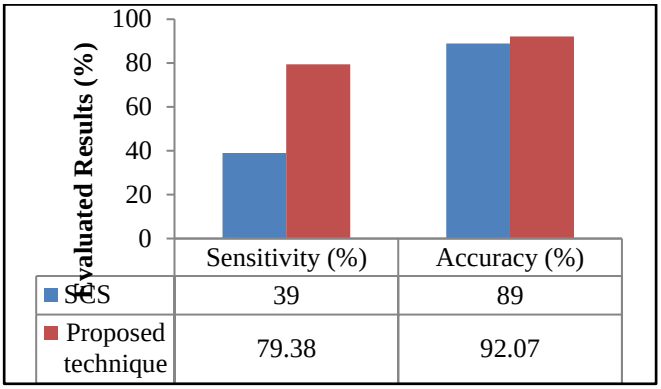

Fig. 4: Comparison for bare category

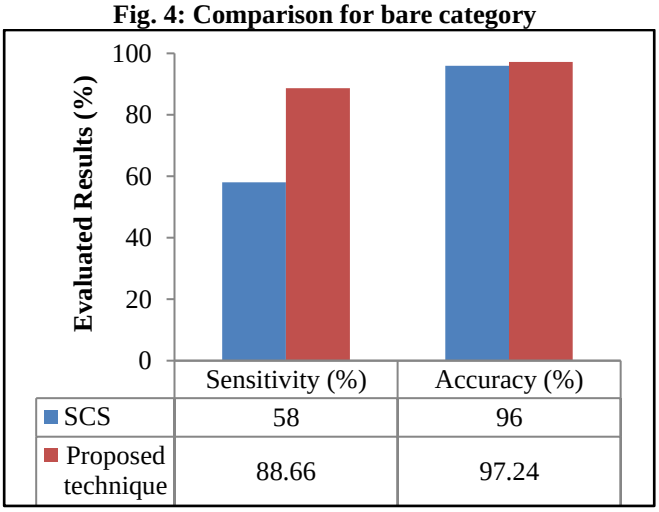

Fig. 5: Comparison for lemm category

The fig. 4 and fig. 5 indicates that boosting approach have achieved 92.07% accuracy in case of bare category and 97.24% accuracy in case of lemm category which is higher as compared to existing accuracy result values of SCS approach for bare and lemm categories respectively. These comparison results indicates the superiority of boosting approach as compared to existing SCS approach for the both the bare and lemm categories of ling spam corpus.

VI. CONCLUSION

Email spam has become a challenge in the field of information sharing and machine learning algorithms are efficiently tackling with this concern. Although the algorithms of machine leaning are placing the crucial role to tackle with email spam, individual algorithms lacks at some points to effectively determine the email spam. In this research wok, ensemble based boosting approach is applied to determine the email spam. The Adaboost booster is used to ensemble the multinomial naïve bayes and decision tree J48 classifiers. The drawback of multinomial naïve bayes algorithm for the classification of contextual emails and decision tree J48 for noisy and long email is overcome by using the properties of other algorithm with the help of Adaboost. The results of the proposed boosting technique are determined using evaluation metrics of sensitivity and accuracy for the ling spam corpus.

Retrieval Number: K13650981119/19©BEIESP
DOI: 10.35940/ijitee.K1365.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The bare and lemm categories of ling spam corpus are considered and evaluated using boosting approach. The proposed boosting approach has achieved the accuracy values of 92.07% and 97.24% for the bare and lemm categories respectively. The sensitivity values of proposed algorithm are also 79.38% and 88.66% for the bare and lemm categories respectively. The evaluated results indicate the outperformed results of proposed boosting technique in comparison with individual multinomial naïve bayes and decision tree J48 classifiers. Moreover, the proposed boosting results are also compared with SCS algorithm which indicates the superiority of proposed boosting technique. For future direction, machine leaning algorithm can be combined with some computational intelligence approach such as swarm intelligence concepts and neural network for the improvements.

REFERENCES

1. Z. L. Bawm and R. P. D. Nath, "A Conceptual Model for effective email marketing", In: Proc. of 17th International Conference on Computer and Information Technology (ICCIT), Dhaka, pp.250-256, 2014.
2. Email Statistics Report, 2018–2022 – Executive Summary, Radicati Group, 2018, Available online at: https://www.radicati.com/wp/wp-content/uploads/2018/01/Email_Statistics_Report,_2018-2022_Executive_Summary.pdf, Last accessed 1 Dec 2018.
3. Talos Intelligence Report, Total Global Email & Spam Volume For December 2018, Available online at: https://www.talosintelligence.com/reputation_center/email_rep, 2018, Last accessed 20 Jan 2018.
4. K. D. Renuka, and P. Visalakshi, "Latent Semantic Indexing Based SVM Model for Email Spam Classification", Journal of Scientific and Industrial Research (JSIR), Vol.73, No.7, pp.437-442, 2014.
5. S. K. Tuteja and N. Bogiri, "Email Spam filtering using BPNN classification algorithm", In: Proc. of International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, pp.915-919, 2016.
6. H. Kaur and A. Sharma, "Improved email spam classification method using integrated particle swarm optimization and decision tree", In: Proc. of 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, pp.516-521, 2016.
7. T. Kumaresan and C. Palanisamy, "E-mail spam classification using S-cuckoo search and support vector machine", International Journal of Bio-Inspired Computation, Vol.9, No.3, pp.142-156, 2017.
8. 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.
9. 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.
10. 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.
11. M. Diale, T. Celik, and C. V. D. Walt, "Unsupervised feature learning for spam email filtering", Computers & Electrical Engineering, Vol. 74, pp.89-104, 2019.
12. I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering", In: Proc. of 11th European Conference on Machine Learning, Barcelona, Spain, pp. 9-17, 2000.

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.

Retrieval Number: K13650981119/19©BEIESP
DOI: 10.35940/ijitee.K1365.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)