Research Article

Email Spam Filtering Using Machine Learning Based Xgboost Classifier Method

[1]p.U. Anitha, [2]dr.C.V. Guru Rao, [3]dr. D. Suresh Babu

[1]PhD Scholar, Dept of Computer Science and Engineering, JNTU, Hyderabad
[2]DIRECTOR, S.R. Engineering College, Warangal, Telangana, India
[3]H.O.D, Department of Computer Science and Engineering, Kakatiya Government College, Hanamkonda
[1]anitha_podishetty@yahoo.co.in,[2]guru_cv_rao@hotmail.com,[3]sureshd123@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: E-mail spam, known as undesirable Bulk E-mail (UBE), junk mail, or undesirable commercial e-mail (UCE), is transferring undesirable e-mail information, usually with business data, in large amounts to a confused set of recipients. Spam is standard on the Internet because automated communications' transaction costs are lower than other alternative forms of communication. Many spam filters use various approaches to recognize the incoming message as spam, varying from white list/blacklist, Bayesian review, keyword matching, postage, mail header analysis, enactment, etc. Even though we are still involved in spam e-mails every day, this paper proposed an enhanced spam exposure design based on Extreme Gradient Boosting (XGBoost) model. It is studied for increased accuracy in spam detection. To the best of our experience, it has expected minor considerations spam e-mail detection difficulties. We explore the proposed system's performance using a more comprehensive range of experimental metrics behind the accuracy, which has managed the existed investigations. The proposed algorithm comparing with existed classifiers of SVM, CNSA-FFO, Rotation forest, MLP, J48, and Naïve Bayes. The proposed model gets better accuracy with 95% when compared with previous classifiers.

Keywords: Email spam filtering, machine learning, XGboosting, Training data, Testing data, Anti-spam.

1. Introduction

The Internet has become an essential part of normality, and email has become an effective tool for recording prospects. Along with the rise of the Internet and email lately, there has been massive spam. Spam can originate from any part of the industry in which you have access to the Internet. Although the advancement of technology and offers of spam, the type of spam is overgrowing. To address the evolution, every organization should look at its system to determine its satisfaction with preventing spam in its environment. Tools, along with a company email and email tool [1].

Portal filtering, getting smaller offers for spam, and human faculty blocking provide an essential arsenal for any business endeavour. However, customers cannot avoid the serious problem of dealing with large amounts of spam every day. Without anti-spam sports, spam will affect societal structures, impair employee productivity, and steal bandwidth, yet it is likely to happen tomorrow.

Spam, known as Unsolicited Bulk Email (UBE), Spam, or Unsolicited Commercial Email (UCE), consists of sending unsolicited email messages, regularly with commercial content, to an organization to an extent. Big. A random number of recipients. The technical definition of spam is "email correspondence is spam" if (a) the non-public identity and recipient context are not relevant because the message applies similarly to many different recipients, and (b) the recipient who can be verified is not given. Intentionally, express and revocable permission to send it. "The danger of spam filtering is that valid email messages can sometimes be rejected or rejected, and valuable email messages can be marked as spam. The ability to stop spam filtering is the regular flow of unwanted mail that clogs networks, negatively affects human inboxes, drains valuable assets such as bandwidth, storage capacity, misplaced productivity, and interferes with the rapid transfer of valid email messages [2].

Spam filters can be implemented at all layers, and firewalls are at the front of the email server or MTA (Mail Transfer Agent). Email Server offers an anti-spam and virus solution that provides complete network-side email security before sending spam or potential email to the network. In the MDA (Mail Delivery Agent) degree, spam filters can also be delivered as carriers to all of your customers. A person can have custom spam filters that automatically filter the email according to the required criteria in an email message, as shown in figure 1. Suggests the same structure as above for the spam filter [3].

2. Problems with spam mail

Cost - The costs of dealing with spam can be prohibitive because spam will be available in large quantities on every customer account, resulting in a loss of network data and bandwidth. It is a significant factor in calculating the total costs of dealing with spam globally [4].

Privacy - Spammers send fake websites that link financial institutions' websites to log in with password and consumer ID or credit card tracking to get consumer credentials and abuse stats. The scam attempts to force the users' account.

Time - Time is the essential element that exists because no one has time to prove that email arrives in the spam or not now. There is no longer enough time to check header stats that identification will come in Mail, in a busy life schedule, Junk Mail It is the main inconvenience to deal with in a long time because it includes a large amount.

Security - In the sheer volume of spam emails, the device's security is compromised because phishing and fraud try to seize customer accreditation records, that is, to open a free account using contact, mobile phone number, and email and aadhar number. And private information such as a debit card [4].

3. Review of literature

There are various implementations suggested by various researchers in email spam filtering using machine learning and deep learning techniques. Some of the literature works described below.

Govil et al. (2020) In the latest internet data is very obviously receiving spam emails. Most of the time, these emails are synthetic. But in many cases, these emails may also include some phishing hyperlinks that contain malware. This arises from the need to introduce a prudent mechanism for accessing or observing these spam emails so that the device's time and memory space can be stored as much as possible. In this document, we introduce the exact mechanism that can filter both spam and non-spam. Our proposed set of rules creates a dictionary and functions and trains them by mastering the machine for practical effects.

Amani Alzahrani et al. (2019) The SMS service became popular after it first became a provider within the second generation's fabric (2G) Land Mobile Community (GSM). Some advertising companies and others have exploited its popularity to spread unwanted advertisements, talk about marketing offers, and send unwanted fabrics to end-of-service customers. These annoying messages called spam, make it difficult for customers to receive perfect notifications and leave them frustrated and angry. Consequently, there are measures taken by many professionals to filter these spam messages and prevent them from reaching terminated customers. Most solutions followed the success of screening unwanted email and used system learning techniques to remove spam messages. A popular device that gains knowledge of techniques that have been effectively used includes Logistic Regression, Support Vector Machine (SVM), Naïve Bayes algorithms, and Neural Networks. The previous study adopted these strategies to filter spam and calculated its accuracy to determine the simplest way to filter out unwanted messages.

Hezha et al (2019) Email is one of the most practical ways to transmit messages between people around the world. Its features, specifically the reliability, speed, and price of the coffee, make it popular and valuable among people in most agencies and society. On the other hand, this recognition has also led to the emergence of new malicious movements, including email (spam) attacks, in our online world. It can be said that spam is one of the main reasons why online strangles with many copies of similar messages created via anonymous senders, which leads to a waste of time/space for the email of the email account holder and also a high risk of viruses and malware for the email providers. Although various filters are used to address the spam issue, including device awareness and content-based filtering, spammers can bypass these protection mechanisms. In this document, we investigate the use of string matching algorithms to detect unwanted email. In particular, this work examines and compares the performance of six known string-matching algorithms, particularly the most extended popular longest common subsequence (LCS), Bi-gram, Levenshtein Distance (LD), Jaro-Winkler, and TFIDF in several combinations. The data is Enron corpora. And the CSDMC2010 Spam Data Collection. We determined that the Bi-gram algorithm reproduces the quality of spam detection in both data sets.

Phaneendra et al. (2019) As the current development of conversation technologies has revolutionized the industry, email has emerged as a widely known conversation model in various business technologies. Email is a powerful, dynamic, and simple messaging method. Statistics sent to emails spam is not required. Spam can be a massive nuisance for both customers and ISPs.

In some cases, spam can also harm the popularity of a specific implementation model. You can see that during most of the more popular mail services, the spam filters bias in choosing the commercial filters. They allow some exceptions for some companies that pay for advertising, marketing and marketing. This isn't always an ethical exercise, but it does pay off in your specific business processes. Our goal is to build a Python hardware-aware spam detector with NLTK, Matplotlib, Word Cloud, Math, Panda and NumPy applications. With this recommended post, we will categorize a specific message as SPAM or SPAM. You can apply using Bayes' theorem, which is easy but effective.

Karthika Renuka et al.(2018) Email is one of the most popular and widely used communication methods due to its worldwide accessibility, high-speed transmission of messages, and coffee delivery charges. Failures of email protocols and the growing number of e-business and financial transactions contribute to the increase in email-based threats. Spam is one of the biggest problems on the internet today, as it causes financial damage to agencies and individual customers. Spam invades customers without their consent and fills up their mail containers. They consume extra community capacity, plus time to scan and delete spam. The vast majority of internet clients openly reject spam, although many respond to commercial propositions that spam remains a vital source of income for spammers. While most users need to think properly to avoid spam, they need clear and simple instructions on how to act. Despite all the steps taken to delay spam, it has not yet been removed. Also, although the countermeasures are very sensitive, even valid emails can be deleted. Among the advanced techniques for preventing spam, one of the most important is filtering. Much of the research on spam filtering has focused on the issues associated with more complex sorting. Currently, machine knowledge of the spam category is a critical research problem. The proposed work's effectiveness is exploring and determining the use of different knowledge algorithms for classifying spam emails. Also, standard metering is provided for several algorithms.

Ghulam Mujtaba et al. (2017) Personal and business customers prefer using email as one of the crucial sources of communication. The use and relevance of email messages is constantly increasing despite the proliferation of alternative forms, including emails, cellular packages, and social media. As the volume of critical business emails continues to grow, the need to automate email management will increase for several reasons, including spam classification, phishing email categorization, and spam categorization. Several volumes, among others. Others. This paper comprehensively criticizes articles on email type published in 2006-2016 using a systematic selection analysis on five elements, in particular, email classification software areas, data sets used in each facility area, and feature area applied at each facility location: technical email classification, and use of performance metrics. A complete evaluation and analysis is carried out to determine the different areas in which the email category is developed to achieve the consultancy's goal. Also, several sets of generic statistics, units of capability, category techniques, and overall performance metrics are tested and used in each specific software area. This overview identifies five application areas for the email category. Information sets, feature sets, classification techniques, and general performance measures are most commonly noted in specific software areas. The extensive use of standard record sets, feature sets, type strategies, and performance measures are stated and justified. Study guidelines, research challenges, and open issues in the email category are also presented to future researchers.

4. Problem defination

Nowadays, email spam is most crucial in social networks. Many problems arise through spam. Spam means that unwanted messages or end-user do not need in their mailbox. Due to this SPAM, the overall system performance can decline and further affects the system's accuracy. Sending unwanted messages which also known as junk mail, are used in spam. At this company, explain spam, as SPAM can damage mail system performance. In the discussion above, there are many types of spam sorters found in spam and non-spam.

Various special email filtering methods are also applied to detect spam. The most famous filters or classifiers are the Decision Tree (DT) Classifier, Naive Bayes (NB), Genetic Algorithm (GA), Support Vector Machine (SVM), etc. We realized that Support Vector Machines are used for the previous testing spam category. But spam takes a long time to be detected. SVM classifier [11] also misclassified messages. So the device may be at risk. The error rate for the SVM classifier can be very high. In this task, there is also a dialog in the feature selection procedure. There are unique job extraction strategies that are used to extract the message

Solution of the problem

To solve the difficulty of existed study in this paper, we are using the XGboosting classifier to classify spam and non-spam emails. The XGboost Classifier is one of the most famous and most accessible techniques for classification. It is a very scalable learning obstacle the number of features needed for the number of linear parameters. It can efficiently train the extensive data quickly with XGoost, which uses less time than another classifier. The accuracy of the system is frequently using this classifier.

5. Methodology

Email spam classification is a big problem in today's computerized environment. A unique spam class method are used to work around this problem. Using this spam detection approach, we can quickly recognise spam and non-junk emails in our mailbox. In this approach, a machine learning-based supervised extreme gradient boosting (XGboost) classifier is proposed to spam email detection and investigated to enhance spam detection accuracy. XGBoost is a perfectly scalable alternative of the Gradient Boost Machine (GBM) classifier that has found full-size advantages in various Instrument Studio competitions. Its management efficiency and its similarity with its great predictive accuracy have made it a target machine for various machine learning (ML) researchers in today's epoch of vital statistics [12]. This method has been used to solve many complex issues in relevant systems, overall performance, and essential generalizability. Some of the significant convenience areas involve infiltration detection, drug discovery, biomedicine, earth, and environmental sciences. Its widespread use, spam detection frameworks performance, and detailed study have yet to be examined to satisfy our experience. Therefore, we analyzed using a source dataset utilized in existing illustrative task under a related experimental service in an extended range of functions. We summarize the proposed work contribution in this research area as given below.

- Enhanced email spam detection method recommended using Xgboost classifier.
- We examine the advanced method performance with earlier actions on the same data
- We examine the proposed method accuracy using a more extended range of evaluation metrics behind the accuracy, which has managed the early investigations

Working of anti-spam

It is completed in two stages: classification and action. In the stage of classification, the message is about whether spam or not. It is also used to describe spam or harmful messages and search for properties to derive them. Once the separation is complete in the transportation section, the message is spam or not, it will be rejected or marked for transfer to the mailbox.

Step − 1: collect the spam and non − spam data

Step − 2: Pre − processing the data means to remove the unused field and

refresh the data

Step − 3: classify the bad sender specific or check the header of mail

Step − 4: apply appropriate classification methods to test according of our requirement

Step − 5: Result stored in text based, image based or content based

The below diagram indicates the anti-spam flowchart

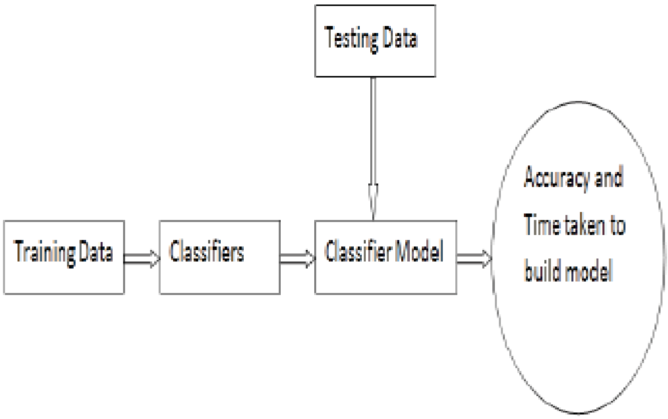

Fig.1 workflow on Anti-spam

Description: In this work, we define the process used to implement a type of spam. The first action is to choose the file from the dataset and follow the selected property's function extraction approach. For it, we use the set of rules for counting sentences. The next step is to educate the dataset to be extracted using job extraction technology. To mark the records, we can calculate the possibility of unwanted and unwanted words inside the file. The further step is to inspect the information with XGBoost Classifier's help to determine spam and nonspam's outcome and predict the higher price. If the spam words are more than the nonspam in an email message, the email is spam mails or nonspam email messages.

Also, we evaluate the words incorrectly categorized with the classifier's advice and determine the classifier's accuracy. It also estimates the classifier's error rate by evaluating the misclassified phrase's fraction and the comprehensive overall quality words in the document.

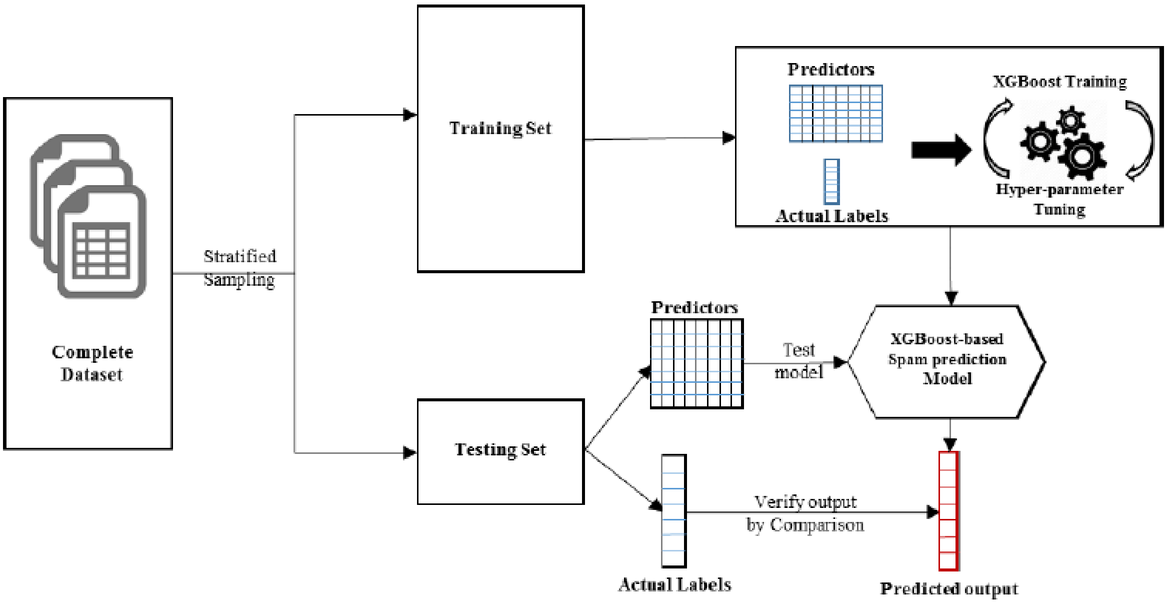

Fig.2 The overall experimental work flow of proposed method

The implementation and experimental task for the proposed XGboost classifier shown in figure 2. The practical data set was split into two sets of training and testing set broad using the sample's stratified sampling method in a ratio with 7: 3, sequentially. While the core of the testing set is 70% showing the training model to the quality and type of spam and non-spam emails, respectively, to test the proposed method overall performance testing set is 30% based on data representations taken.

The proposed model design was taken out in interpreted and objected-oriented python programming language with the XGBoost package used in the modelling process.

The supervised learning-based XGBoost method is briefly explained in below.

The suggested XGBoost method for spam classification is supervised learning-based machine learning, which is usually an aggregate of a set of weak regression Tree and classification problems, say $k, \{T_1(x_i, y_i) \dots . T_k(x_i, y_i)\}$. Meanwhile, regression Tree and classification (RTAC) is an individual decision tree that connects an actual count to every result. The prediction score of every weak tree is added up, and the resulted score calculated using K additive functions and f_k is space of all conceivable RTAC as equation for the F shown below

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (1)$$

The daily goal function is given by equation 2. The initial and second sentences are characteristic of differentiable loss - the degree of distinction between $\widehat{y_i}$ and the expected target y_i,, and the time of agreement - broadcast a degree of complexity.

$$Obj(\theta) = \sum_{i}^{n}(y_i, \widehat{y_i}) + \sum_{k}^{K} \Omega(f_k) \quad (2)$$

The regularization duration is given by $\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j-1}^{T} w_j^2$ (3)

Where the vector of counts on each of the leaves and the no. of leaves are indicated by w and T separately. The further constant $\gamma \ and \ \lambda$ are used to manage the regularization degree. Additional techniques utilized to avoid overfitting in XGBoost are data subsampling and shrinkage.

The additional training used in XGBoost indicates that the prediction $\widehat{y_i}$ appears at t that can expressed as equation 4.

The results objective property of the $t - th$ tree appears afterward attractive the Tailor's extension of loss function and increasing the term regularization as given in equation 5

$$\widehat{y_i}^{(t)} = \sum_{k=1}^{K} f_k(x_i) = \widehat{y_i}^{(t-1)} + f_t(x_i) \quad (4)$$

$$Obj\ (\theta)^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}\ h_i f_t(x_i)^2\right] + \gamma T + \frac{1}{2}\lambda \sum_{j-1}^{T} w_j^2 \qquad (5)$$

$$= \sum_{j=1}^{T}[\left(\sum_{i\epsilon I_j} g_i\right) w_j + \frac{1}{2}\ (\sum_{i\epsilon I_j} h_i + \lambda)\, w_j^2] + \gamma T \qquad (6)$$

Where,

$I_j = \{i|q(x_i) = j\}$ is instance set of leaf j. For a $q(x)$ $tree\ structure$, w_j^* is the weight of optimal leaf and the objective functions are attained using below equations 7 and 8.

$$w_j^* = -\frac{G_j}{H_j+\lambda} \qquad (7)$$

$$Obj^* = -\frac{1}{2}\sum_{J=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T \qquad (8)$$

Where

$G_j = \sum_{i\epsilon I_j} g_i$ and $H_j = \sum_{i\epsilon I_j} h_i$ Later, every leaf is divided into two score intended again with equation 10 as shown below

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} + \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \qquad (9)$$

Where the first 3 sentences indicates the degrees of the left, right and original leaf correspondingly, while the remaining term is the organization in the extra leaf. The summarization of at XGBoost is as follows;

Algorithm: XGboost for email spam filtering

```
Step1: Select the file
Step2: Extracting the feature with the help of word count algorithm
Step3: Training the dataset with the help of XGboost classifier
Step4: Detect the probability of spam and non − spam mails.
Prob_spam = (sum(train_matrix(spam_indices, )) + 1))/(spam_wc + numtokens)
Prob_nonspam = (sum(train_matrix(nonspam_indices, ))
+1/(nonspam_wc + numtokens)
Step 5: Testing the dataset
log_a = test_matrix ∗ (log(prob_tokens_spam))' + log(prob_spam)
log_b = test_matrix ∗ (log(prob_tokens_nonspam))' + log(1 − prob_spam)
if
output = log_a > log_b
then document are spam
else
the document are non_spam
Step 6: Classify the spam and non spam mail
Step 7: compute the error of the text data and calculate the word which is wrongly classified
Numdocs_wrong = sum(xor(output, text_lables))
Step 8: display the error rate of text data and calculate the fraction of wrongly
classified word
Fraction_wrong = numdocs_wrong/numtest_docs
```

6. Dataset desciption

In this paper, we are using 'spam_ham_dataset' from kaggle and the dataset includes total 5172 records with 4 columns. In the dataset email was considered spam (1) or not (0) i.e. undesirable business e-mail. The Dataset divided into two categories 70% is training set and 30% is testing set. The experimental conducted on the needed dataset of both training set and testing set.

Evaluation metrics

To confirm impartial comparison with existing works, the following are frequently utilized metrics to get the performance of the suggested method.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (11)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (12)$$

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN} \qquad (13)$$

$$Precision = \frac{TP}{TP+FP} \qquad (14)$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (15)$$

Where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative respectively

Expermental results

The proposed model experimental results are evaluated and calculated using various precision, Recall (Sensitivity), specificity, f1-score, and Accuracy metrics. And the practical conducted on both training and testing datasets that values are shown in the table.1.

Table.1 Test results in percentage

Data	Sensitivity/Recall	Specificity	Precision	F1-Score	Accuracy	ROC-AUC	PR-AUC
Testing/training	95.59	97.73	96.47	96.03	95	99.08	97.69

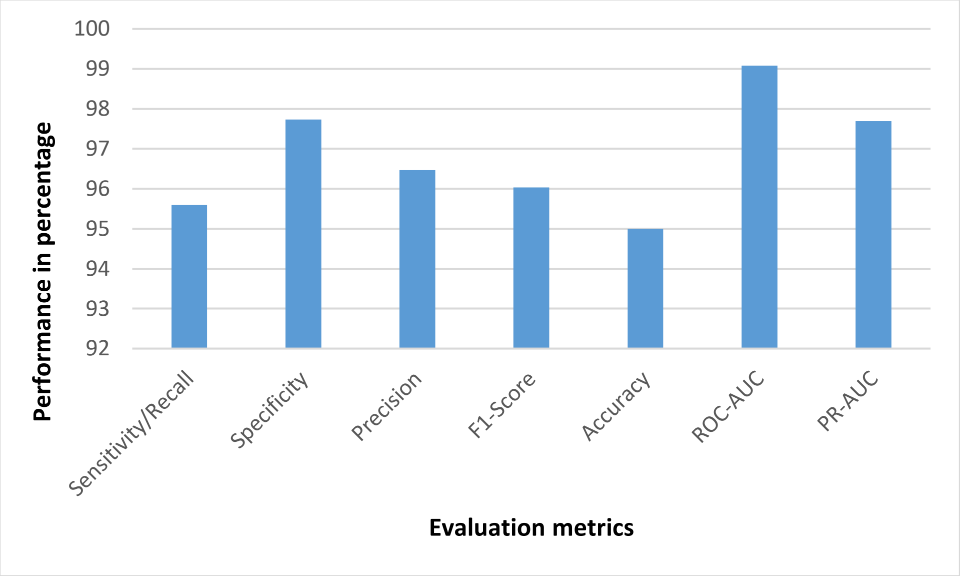

Fig.3 Proposed method performance calculation

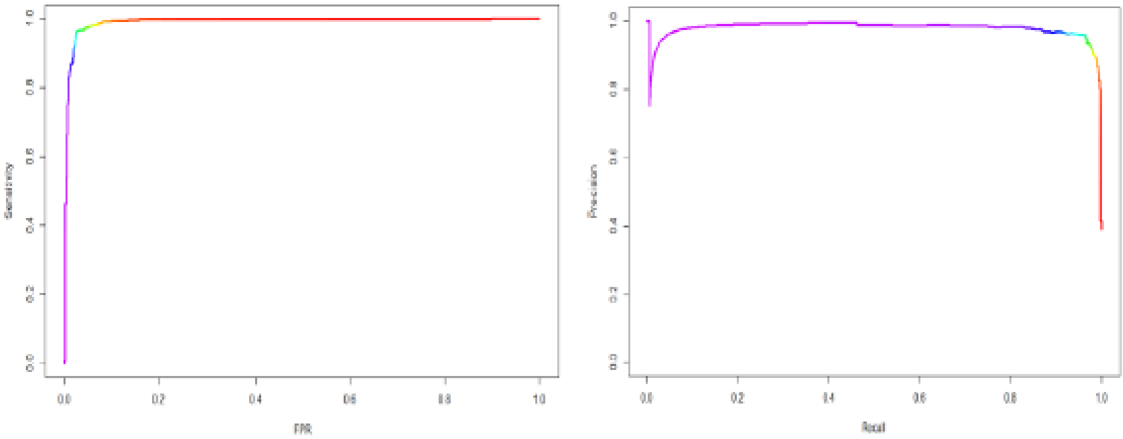

Fig.4 (a)ROC and (b) PR Curves

Table 2: comparison between various methods testing performance with proposed XGboost classifier

Classifier	Accuracy	Sensitivity	Specificity	Precision	F1-Score	ROC-AUC
Proposed XGBoost	95.00	95.59	97.73	96.47	96.03	99.08
SVM[11]	94.06	93.87	94.06	-	-	-
CNSA-FFO[13]	93.88	87.28	97.31	-	-	-
Rotation Forest[14]	93.50	93.50	-	93.50	93.50	97.60
J48[14]	91.20	91.20	-	91.20	91.10	93.70
MLP[14]	92.30	92.30	-	92.30	92.30	97.30
Naïve Bayes[15]	91.13	86	-	88	77	

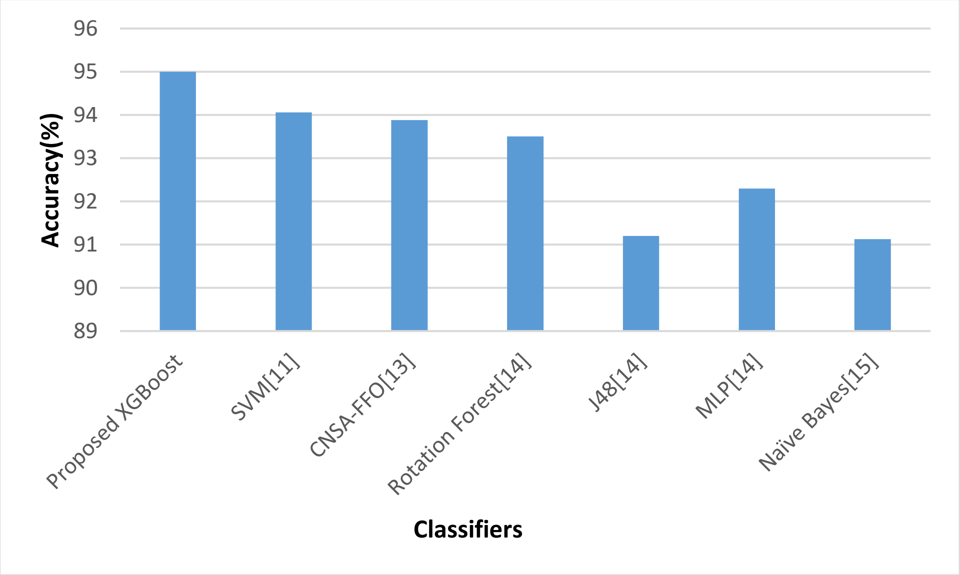

Fig.5 Accuracy comparison between various classifiers

As shown in figure 5, the accuracy comparison is taken between various classifiers. In the figure, the X-axis indicates the classifiers, and the Y-axis indicates the accuracy (%), and the proposed model provides better accuracy with 95% compared with other classifiers.

7. Conclusion

Email spam is one of the most crucial in today's world. To solve this issue, various authors proposed many automated classification techniques to find spam or nonspam emails. Here we are used the XGBoot classifier to detect the spam emails from that given dataset. The proposed XGBoost classifier used as a spam emails indicator and utilizes a standard computational intellect method on a benchmark dataset. Empirical outcomes are a valuable classifier with optimized hyper-parameters that is less sensitive to overfitting. The proposed method compared with current classifiers of SVM, CNSA-FFO, Rotation forest, MLP, J48, and Naïve Bayes classifiers. The evaluation results confirm that the proposed model got better accuracy with 95% compared with the current approaches.

References

1. STeli Savita and Biradar Santosh Kumar, 2014, "Effective Email Classification for Spam and Non-spam," pp. 273-278.
2. S Rushdi and M Robet, 2013, "Classi_cation spam emails using text and readability features", IEEE
3. Aditya Shrivastava, Dr. Rachana Dubey, 2018, "Classification of Spam Mail using different machine learning algorithms", IEEE.

4. Prof. R.S and Ms. Rachana Mishra, 2013, “Thakur: Analysis of Random Forest and Naïve Bayes for Spam Mail using Feature Selection Categorization”, pp. 43 – 48.
5. N. Govil and Astha Varshney, 2020, “A Machine Learning based Spam Detection Mechanism”,IEEE, pp.954-957.
6. B. Rawat Danda and Amani Alzahrani, 2019, “Comparative Study of Machine Learning Algorithms for SMS Spam Detection”,
7. Hezha M. Tareq Abdulhadi and Cihan Varol, 2019, “Comparison of String Matching Algorithms on Spam Email Detection”, pp.6-11.
8. L. Phaneendra M, R. Ragupathy, 2019, “Adaptive Prediction of Spam Emails: Using Bayesian Inference”, pp.628-632.
9. D. Karthika Renuka, Lakshmi Surya P, 2018, “Spam Classification Based on Supervised Learning Using Machine Learning Techniques”
10. Ghulam Mujtaba, N.Majeed, 2017, “Email Classification Research Trends: Review and Open Issues”, pp.1-21.
11. S.O Olatunji, 2017, “Improved email spam detection model based on support vector machines”, pp. 1-9.
12. Nielsen, D., Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition? 2016, NTNU
13. S. Chikhi and R Chikh, 2019, “Clustered negative selection algorithm and fruit fly optimization for email spam detection”, pp. 143-152.
14. M. Shuaib, et al, 2018, “Comparative Analysis of Classification Algorithms for Email Spam Detection”, pp. 60.
15. Nurul Fitriah Rusland, Hanayanti Hafit, 2017, “Analysis of Naive Bayes Algorithm for Email Spam Filtering across Multiple Datasets”, pp.1-9.