Analysis of Machine Learning Approach for Spamming Electronic Mail Detection

Sajan Shrestha1, Parth Jain2, Gururaj H L2, Soundarya B C2, Gurushankar H B2

1Dept. Of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Bengaluru, India

2Dept. of Information Techonolgy, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Bengaluru, India

**Abstract.** Electronic mail (email) is a standard way of communicating with peoples but an increasing number of email users have increased different unethical activities like irrelevant or unsolicited messages, sending malwares, etc. Such emails that contains mentioned features are commonly called as spam emails. Machine learning algorithm such as Logistics regression, decision tree, random forest, SVM, Navie Bayes and BERT are used for the detection of spam emails.

In this research work, data containing various features of emails such as subject of email and message in it are vectorized into spare matrix using feature extraction(TFIDF) for logistic regression, decision tree, random forest,SVM, Navie Bayes.

Used BertTrasnformer to convert the text into the token which will be used as input for BERT.

Considering multiple parameters, Random Forest and Decision Tree Algorithms achieves 98% accuracy

**Keywords:** Cybersecurity, Spam, attack, Machine Learning,BERT

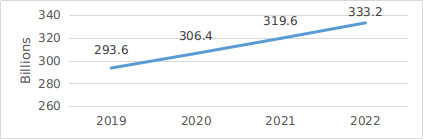
1. INTRODUCTION

Everyone are using emails for sending official letters and connecting with the peoples all over the world. Such expansion of this technology has benefited many but along with its development, it has created major problems like fraud text, transferring virus through file sharing and irrelevant or unsolicited messages. It is clear that spam emails contains text bugs and programmed virus as depicted in Fig. 1. Those mails are considerably known as spam mails and the process of searching those spam mails is known as spam filtering.



**Fig. 1.** Spam Mail

In 2019, daily spam number was 293.6 billions which has increased to 333.2 billions (48.63 % of total emails[16]) in 2022[17]. According to the data represent in Fig. 2,the number of spam emails per day is increasing in a positive linear fashion. If it follows same trends then by 2025 there will be around 400 billions of spam emails per day.



**Fig. 2.** Email spam volume per day

Various organization and researchers are spending billions of dollors for finding the most efficient algorithms which can detect the spam emails. Evolution of Artificial Intelligent(AI) and Machine learning(ML) have made the detection of those spam emails lot more easier. Some popular machine learning algorithms that are used for email detection are Logistic Regression, Support Vector Machine(SVM), Naive Bayes, Decision Tree and Random Forest. Despite such improvement in AI and ML, 100% efficiency cannot be achievable unless the model is trained with numerous features such as domain reputation, url present in emails, user feedback for that email address and many more.

In this research work, various machine learning algorithms are used for email spam detection. The main focus is on data pre-processing where all the additional features (sender reputation, phishing websites, header analysis) of emails are included for finding whether the corresponding email is spam or ham. All those additional features are vectorized using TF-IDF for modeling in different machine learning algorithms. The result obtain from different machine learning model are compare based on confusion matrix and classification report which includes accuracy, recall, precision , F1-score.

The paper is structured as follows: Section-2 explains the related work of the existing works. Section-3 elaborates the methodology. Section-4 delivers the results and finally conclusion has been drawn in section-5.

1. LITERATURE REVIEW

In this section, related work has been depicted with existing works.

Billions of peoples are using emails and the major problem they are facing is inevitable spam emails which contains malwares and wrong information. This results loss of time and money and it is estimated that by 2025 there will be loss of $10.5 trillions because of spam emails[17]. Such spams can be detected using spam filter which differentiate the spam emails with ham emails based on the content of emails, email header details, the route that it takes from its original sender.

Some major types of spam filter for email spam detection are given below[6]:

### *Bayesian Filters.* These filter is based on Bayes theorem which determines the probability that a message is spam or ham based on the words and phrases it contains.

### *Content-Based Filters.* Content-based filters examine the actual content of messages for certain keywords, phrases, or patterns commonly associated with spam.

### *Header Analysis Filters.* Header Analysis filters analysis the email header information like sender’s email address, IP address, subject line etc to distinguish between spam and ham emails.

### *Machine learning Filters*. In this filters, different Machine learning algorithms are used for detection of spam mails. Those algorithms includes decision tree, random forest, logistic regression, neural networks, support vectors etc.Machine learning filters is trained using data set and along with the time the accuracy of this filters can be improved.

### *Sender Policy Framework(SPF).* SPF is an email authentication protocol that helps verify the legitimacy of the sender's domain. Spam filters can check SPF records to determine if an email is coming from an authorized source.

Although there are many other spam filters, the spam traps will still be there. So, many organizations and researchers are aiming to build robust and efficient filters to stop this problem. Many models based on machine learning algorithms show excellent performance to detect whether an email is spam or ham. Many researchers have got above 90% accuracy in their related algorithms along with that they have consider features like email text, header information, DNS record.They have used popular machine learning algorithms and Deep learning algorithms for training their model.

Spam email can also be detected based on the link present on it. For example, if mail contains phishing websites then it is obvious that the email is a spam. Detecting a phishing websites can be determined based on the features like address bar, abnormality, HTML and java script and Domain information.[5].

Saleh et al.[18] presents spam filters that is based on Negative Selection Algorithms(NSA) that can detect its own type(normal or self pattern) but not the new types(different pattern). In this study, Negative Selection works same as T-cells in biological system as it recognize its own kind(normal pattern) but attack other external entities(different pattern) .Ahmed, Naeem Ahmed [3] uses machine learning techinque for spam filters and classify different types of spam filter (standard spam filter, client side spam filter, enterprise level spam filter and case based spam filter).

Thashina Sultana [1] uses email spam detection using machine learning approach (Naive Bayes) using IP address and email text as parameter. Similarly, focusing on data preprocessing procedure, Nikhil Kumar [2] used different steps like data cleaning, data integration,data transformation and data reduction. Yaseen, Qussai [8], concludes that spam email wastes millions of dollars annually. In this research paper, they uses Deep neural network model for such spam detection. Kartik [7] uses machine learning approach and For result comparison kappa coefficient, matthew correlation coefficients are used in this research work. Debnath, Kingshuk, and Nirmalya Kar [4] have used enron email data set in deep learning models like LSTM and BERT. NLP was used to analyze data pre processing of text of email.

This research paper consider additional features of emails which are vectorized to spare matrix using feature extraction and uses that spare matrix to train and test the machine learning algorithms model. Moreover, this paper includes the comparative analysis of efficiency of model which is trained with additional features and without additional features.

1. METHODOLOGY

In this section the detailed methodology has been explained which includes data collection , data pre processing, modeling and testing techniques.

## 3.1. Data Collection

Data is collected from the Kaggle.

Link: <https://www.kaggle.com/datasets/shantanudhakadd/email-spam-detection-dataset-classification>

Parameters: message, spam or ham

## 3.2. Data Pre-processing

Data preprocessing is the procedure of converting the raw data into a vectorized dataset which can be used for training and testing learning models. The first step for data preprocessing is data cleaning which corrects the missing values, removes inconvenient data entry and modifies the data set as per the condition of the dataset[11].

### **TFIDF.** TFIDF stands for term frequency- inverse document frequency, is a numerical value that defines the significant of a word in given document. The numerical signigicant of a word is calculated using vectorization and feature extraction in which TF-TDF is calculated for each word using formula as shown in Equation (1) .

Using this TF-IDF, a sparse matrix is calculated which will have n row (distinct types of word in the document) and m column (total number of text)[12]. In addition, we add more columns to this sparse matrix which give the other information about email like sender reputation, phishing website etc. For example,additional column is made for phishing websites in which the range of given phishing website is calculated: the most popular phishing website, the value is high and safe one, the value is almost zero.

The mathematical expression for TF-IDF is given below[19]:

*TF-IDF:W(i,j)=F(i,j) x log(N/F(i))*  (1)

where,

W(i,j): Word i in document j

F(i,j): Frequency of i in documentj

N: Total number of document

F(i): Total frequency of word i in all document

## 3.3. Modeling

Machine learning model is made for spam detection which separate ham from spam emails based on the information about that email. Making a model for a testing purpose is done by using different supervised machine learning algorithms.

### **Logistic regression.** In initial phase of logistic regression, random value or zero value is assign to the parameter θ and by looping the Equation(2) formula of gradient descent, the optimum value of parameter θ is determined. This overall method of finding the parameter is known as modeling. Furthermore, the determined optimum value of parameter θ is used to determine value of Sigmoid (Logistic) function using Equation(3).It provides the value between 0 to 1[20]. If it is less then 0.5 then it is consider as ham otherwise it is spam.

For modeling, optimal value of θ is calculated using gradient descent :

θ=θ-Σ[h(x)-y].x (2)

*h(x)=* : Sigmoid function (3)

where,

m is the total number of features

Y: 0: Spam and 1:Ham

X: Features of email

### **Naïve-bayes.** This algorithm is based on Bayes Theorem which gives the conditional probability of an event A given that another event B had occurred[13]. Using the Equation(4), the probability of spam email is determined with the help of data present on it.

*P(A/B) =*  (4)

where,

P(A/B): Conditional probability of A given B

P(B/A): Conditional probability of B given A

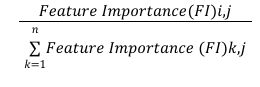
P(A): Probability of event A

P(B): Probability of event B

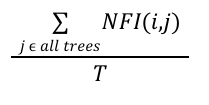
### **SVM.** SVM stands for support vector machine. In this algorithm, a margin line is determined to distinguish data points between two different classes. In addition, two extreme hyper planes (positive and negative) are also determined to separate the classes. In SVM, kernel function is applied in a given data set to manipulate data set which makes it easier to find the margin line for that given data set.

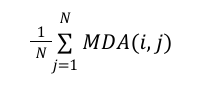
### **Random Forest.** Random forests (RF) construct many individual decision trees at training and the result of random forest is calculated by averaging all the predicted values of all decision trees. Following equation(5-7) are used to find the final output.

Random Forest Feature Importance

RFFI(i)= (5)

Normalized Feature Importance

 NFI(i,j)=  (6)

Feature Importance

FI(i)= (7)

Where,

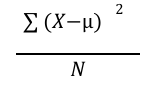
NFI(i,j): normalized feature importance for feature i in tree j.

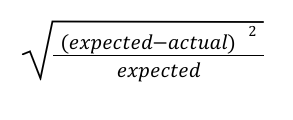
FI(i,j): original feature importance for feature i in tree j.

N,T: Total number of Decision Tree in Random Forest

*MDA(Mean Decrease in Accuracy).* MDA value for specific feature is the amount of decrease in accuracy of model when the value of that feature are randomly permuted while keeping other variables unchanged.

### **Decision Tree.** Decision tree split a data set into two groups based on different features as depicted in Figure 3. In decision tree algorithm, various parameter like gini impurity, varience, chi-square and entropy are calculated using the following Equation (8) and(9).

Varience= (8)



Chi-square= (9)

X: actual value of element

μ: mean value of element

N: Number of element

# 

**Fig. 3 .** Decision Tree

## 3.4. Testing

Testing a model will help to determine whether the algorithm is working efficiently or not.

### **Confusion Matrix.** Confusion matrix is a tool used in the field of machine learning and statistics to visualize and evaluate the performance of a classification model. A confusion matrix is usually a square matrix with rows and columns corresponding to the true classes and predicted classes, respectively.

### **Classification Report.** Classification report is the report which evaluated the efficiency of machine learning algorithms [9]. In classification report there are mainly four parameter which are explained below:

#### *Accuracy.* In machine learning, accuracy refers to the measure of how well a classification model correctly predicts the class labels of the input data points. It is a commonly used metric to evaluate the performance of classification algorithms which is determined using Equation(10).

Accuracy: : (10)

#### *Recall.* Recall, also known as sensitivity or true positive rate,is a metric that calculated using Equation(11) and quantifies the ability of a classification model to correctly identify all positive instances from the total actual positive instances in a data set.

Recall: (11)

#### *Precision.* Precision is tool that measures the accuracy of positive predictions made by a classification model.Precision is the ratio of True Positive and sum of True Positive and False Positive as in Equation(12).

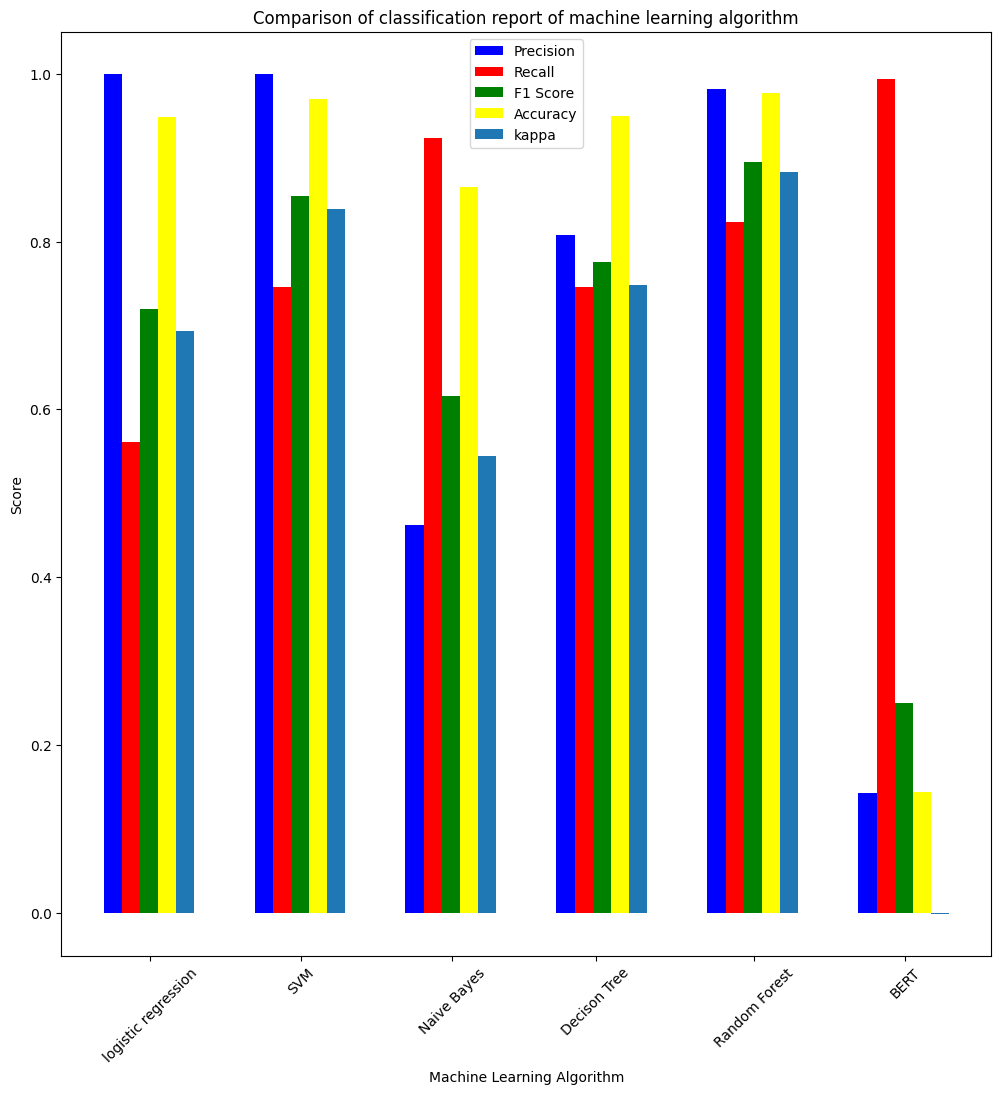
*Precision:*  (12)

#### *F1-Score.* F1 score is a metric that combines both precision and recall into a single valued as shown in Equation(13), providing a balanced measure of a model's performance, particularly in situations where class imbalance exists. classification report of all the machine learning algorithm

*F1-Score: (13)*

1. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the results obtained after testing all the machine learning algorithm are present in tabulation form. Machine learning algorithms like Logistic Regression, Naive Bayes, Support Vector Machine(SVM), Decision Tree and Random Forest are used as a model for training and the comparative analysis of efficiency of those algorithms are tabulated as shown in Table 1.



**Fig. 4.** Spam Classification report without additional features

**Table 1.** Factors and their associated advantages that have not been

considered in market

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metrics: | | Precision | Recall | F1-Measure | Accuracy |
| Logistic Regression | Ham  Spam | 0.93  0.99 | 1.00  0.54 | 0.96  0.70 | 0.98 |
| SVM | Ham  Spam | 0.96  0.99 | 1.00  0.75 | 0.98  0.85 | 0.99 |
| Decision Tree | Ham  Spam | 0.97  0.91 | 0.99  0.78 | 0.98  0.84 | 0.96 |
| Random Forest | Ham  spam | 0.97  0.98 | 1.00  0.83 | 0.99  0.90 | 0.96 |
| Naive Bayes | Ham  Spam | 0.98  0.47 | 0.84  0.90 | 0.91  0.62 | 0.85 |
|

1. CONCLUSION

Email has been the most popular communication media all over the world. Billions of data are transferred through emails daily. Such a vast increasing number of email users has increased the number of spam emails. Spam emails are those emails which contain malware in the form of files, unsolicited messages and many unethical works. This problem initiate spreading of wrong information, malwares through emails that results waste of billions of memories. Spam emails generated from various websites or botnets can be detected using the different features of coming emails at every gateway using machine learning algorithms.

**References**

1. Sultana, Thashina, et al. "Email based Spam Detection." International Journal of Engineering Research & Technology (IJERT) (2020).
2. Kumar, Nikhil, and Sanket Sonowal. "Email spam detection using machine learning algorithms." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2020.
3. Ahmed, Naeem, et al. "Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges." Security and Communication Networks 2022 (2022): 1-19.
4. Debnath, Kingshuk, and Nirmalya Kar. "Email spam detection using deep learning approach." *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*. Vol. 1. IEEE, 2022.
5. Harinahalli Lokesh, Gururaj, and Goutham BoreGowda. "Phishing website detection based on effective machine learning approach." *Journal of Cyber Security Technology* 5.1 (2021): 1-14.
6. Ani Petrosyan, & 8, M. (2023, March 8). *Spam e-mail traffic share 2022*. Statista. https://www.statista.com/statistics/420400/spam-email-traffic-share-annual
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
8. Kartik Ahluwalia, Gururaj HL, Rahsmi R,"Comparative Analysis of Various SMS Spam Detection Methods Using Machine Learning.",2023.
9. Yaseen, Qussai. "Spam email detection using deep learning techniques." *Procedia Computer Science* 184 (2021): 853-858.
10. Kaur, Simarjeet, Meenakshi Bansal, and Ashok Kumar Bathla. "A Comparitive Study of E-Mail Spam Detection using Various Machine Learning Techniques." *AIJR Proceedings* (2021): 426-435.
11. Delany, Sarah Jane, Mark Buckley, and Derek Greene. "SMS spam filtering: Methods and data." *Expert Systems with Applications* 39.10 (2012): 9899-9908.
12. Chandrasekar, Priyanga, and Kai Qian. "The impact of data preprocessing on the performance of a naive bayes classifier." *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 2. IEEE, 2016.
13. Zhou, Hai. "Research of text classification based on TF-IDF and CNN-LSTM." *Journal of Physics: Conference Series*. Vol. 2171. No. 1. IOP Publishing, 2022.
14. Ma, Thae Ma, Kunihito Yamamori, and Aye Thida. "A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification." 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE). IEEE, 2020.
15. Kudupudi, Nikhil, and Shilpa Nair. "Spam message detection using logistic regression." International Journal of Advanced Computer Science and Applications 9.9 (2021): 815-818.
16. *Kaspersky’s 2022 spam and phishingreport*. Securelist English Global securelistcom.
17. Khodun, Y. (2023, May 25). *How manyemails are sent per day?*. MacKeeper. Ani Petrosyan, & 8, M. (2023, March 8).*Spam e-mail traffic share 2022*.Statista.
18. A. J. Saleh, A. Karim, B. Shanmugam et al., “An intelligent spam detection model based on artificial immune system,” *Information*, vol. 10, no. 6, p. 209, 2019.
19. Singh, Ksh Nareshkumar, et al. "A novel approach for dimension reduction using word embedding: An enhanced text classification approach." *International Journal of Information Management Data Insights* 2.1 (2022): 100061.
20. Kumar, K. Varun, and M. Ramamoorthy. "Machine Learning-based spam detection using Naïve Bayes Classifier in comparison with Logistic Regression for improving accuracy." *Journal of Pharmaceutical Negative Results* (2022): 548-554