2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) | 979-8-3503-4314-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSITSS60515.2023.10334171

Efficient Spam Email Classification using Machine Learning Algorithms

Pallavi N
MTech in CNE
Dept. of ISE
B.M.S. College of Engineering
Bangalore, India
Pallavin.scn21@bmsce.ac.in

Dr P Jayarekha
Professor and HOD
Dept. of ISE
BMS College of Engineering
Bangalore, Inida
jayarekha.ise@bmsce.ac.in

***Abstract*— In today's digital age, since email is the main form of communication, the identification of email spam is a critical issue. In addition to consuming a lot of time and money, email spam is also a security and privacy risk. In this paper, we provide a means for email spam detection that employes machine learning Algorithms. The required features for training the ML models have been engineered after analysis of the email dataset of content-based filtering obtained from Kaggle website. We tested a Several types of algorithms for machine learning and analyzed their level of performance using the dataset. Our findings demonstrate how effective is the suggested approach in identifying email spam with highest accuracy of 99.8% and Rmse of 0.2 .Here we applied , the various ML classifier algorithm such as Decision tree , Voting Classifier , Random Forest, Logistic Regression and so on to our dataset ,compared among each other and found which suits best for the dataset with the highest accuracy. This method can be useful in email clients or servers to detect spam emails automatically and enhanced**

Keywords*— *Spam detection , Machine learning, classifier Algorithms , Accuracy

I. INTRODUCTION

In present world , using the email in order to access certain benefits/Features of the online resources. E-mail is a term that millions of people use on a daily basis. To join an online class platform, to transact official Documents or the paperwork, to use it for various banking transactions ,or even to engage in the increasingly popular online shopping, internet users must have an email address registered. To verify their identity and to register on any site, users must need an email address. The use of email has been increasingly dependent on people over the past few months. Daily email sends and receives were 293.6 billion in only 2019 alone.

Nearly 3.930 billion people used email globally in 2019, 4.037 billion in 2020, and 4.147 billion people are using it as of 2021. People of all ages and social classes, including businesspeople, students, and others, are represented in this user list. The number of human traps, or, to say it another way, the hacking and extraction of personal information via e-mail, is dramatically rising along with the user rate. Spam email refers to unsolicited email messages that are delivered to a large list of recipients. It might be used for fraudulent or commercial purposes.

Most of the time, based on the email's subject line, it is can be determined whether it contains harmful material, but the guess or forecast may or may not be correct. Filtering spam emails to identify them and stop the theft of personal information becomes crucial. Both machine learning and non-machine learning techniques can be used to filter spam emails. It should be noted for a comparative comparison that non-ML models deal with deterministic techniques, whereas ML models primarily deal with concepts of probabilistic approaches.

In basic terms, ML models lack rigid rules, but non-ML models do. For this purpose, it is possible to anticipate whether a received email contains malware content using a variety of machine learning methods. And possible to assess the precision of those predictions. This can aid in the detection of spam mail, which can save us and ensure the protection of our personal data. This paper will analyse the dataset made up of spam and junk mail, and then use several machine learning methods to forecast and assess accuracy. If widely used, it can be shown to be useful for identifying the likelihood of malware attacks in received mail.

In This Paper , The existing/known dataset is considered for the analyses,the comparative analysis of several Machine learning algorithms with their Accuracy, RMSE, MAE, MSE, R-squared and Confusion matrix demonstrated visually using the Heatmap and with the Plots and graphs.

II. REVIEW OF LITERATURE

The major goal of a literature review is to analyse the project's background, which reveals any drawbacks in the current system and directs us towards any unresolved issues that need to be resolved.

Datasets are compiled with information from communications. The datasets, which also include the ham

and spam sets, are taken from various sources and saved in an excel file. investigating the word usage in spam and non-spam texts[1]. A pre-existing dataset from Kaggle.com is used for the study[2][4]."Hard ham" appears in the subject line of multiple emails, includes unexpected inputs, such as null values, was challenging to preserve such inputs in a specific pattern or class [2] The second dataset is the Email Spam dataset from the UCI Machine Learning library[4][6], 5674 emails are totaled throughout 5674 rows and 2 columns, each email contains two columns labelled "Category" and "Message," respectively where every communication, including emails, is classed as either spam or not [4].

For the analysing the dataset , The system's accuracy is improved while utilizing the NB Classifier[1]. The dataset was analysed utilising the Python programming language. The "Latin-1" encoding to encode the string as a series of bytes. For ease of understanding, we changed the terms "spam" and "ham" to binary values 1 and 0, respectively.We used the "Bag of Words (BoW)" text modelling approach for NLP[2] The study is centred on evaluating the outcomes of Bayesian filtering and contrasting those outcomes with those of other machine learning methods. The optimal method was ultimately chosen based on the algorithm's accuracy. The naive Bayes algorithm scored highly in a research comparing several other ML algorithms[3]. The input dataset is made up entirely of unprocessed, raw data. The accuracy ratings of each classifier are compared, and For comparisons, the confusion matrix is used. Reports are created for classification.Ensembleclassifier performs better with larg edimensional datasets when compared to other classifiers[4].Examines a range of non-machine learning based frameworks to further emphasise the flaws and the current state of affairs[5]. The detection model was created by converting the text input into vectors, creating a BiLSTM model, and fitting the model with the vectors[6]. Based on the solution likelihood, the result indicates whether or not the message is considered spam. The common keyword visualization is shown with the phrases spam and ham[1]The problem of Hardham due to unnecessary characteristics in dataset is resolved via Feature extraction,The results and model accuracies were achieved after parameter optimisation using the cross validation procedure with five folds , With just the information gain matrix, and categorised the text with an accuracy of 98.445%[2].It accurately predicts the subject line of a bran d-new email[4]. The main topic was automatic, intelligent defence against affecting spam emails ,cover special focus on malicious links, malicious attachments, phishing attempts, spoofing, and phishing tactics also excludes the studies that just focus on spammy marketing emails[5] spam detection model was created by converting the text input into vectors, creating a BiLSTM model, and fitting the model with the vectors. Model with 95% of F1 score with BERT, BiLSTM, and LSTM, the suggested deep learning approach achieved a maximum accuracy of 99.14%, 98.34%, and 97.15 %[6] Based on the specified standards for evaluating the quality of the literature, 52 primary studies from the formal and grey literature were chosen. Phishing, scamming, spamming, smishing, and vishing were the main social engineering-based techniques used during the COVID-19 pandemic, along with the most popular socio-technical technique: the use of fraudulent emails, websites, and mobile apps as platforms for successful cyberattacks. Ransomware, trojans, and bots are three types of malicious software that are routinely used to attack systems and resources. Also highlighted the financial impact of these cyber-attacks on various organisations and critical infrastructure[7].

From the review of the literature , the ML and DL is used for analysis resulting with the less Accuracy and has the drawback of choosing the dataset, as they are not applicable for every other data of different pattern.To overcome the problem here we provide the method for the data on content based filtering,Blacklisting and whitelisting.

III. METHODOLOGY

I.

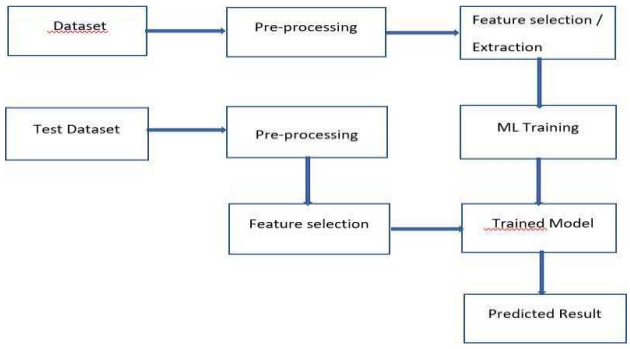

Figure. 1 System architecture

The above figure represents system architecture of proposed system, where we are applying various ML algorithms to classify the Email spam It consists of :

- Dataset / Testing Dataset
- Pre-processing
- Feature Extraction
- ML Training
- Training Model
- Predicted result

In this part we discuss about the methods we used in each stepof the analysis of Email spam classification

I. Collection of dataset ;
The Existing dataset is considered for the study.The data is been obtained from the Kaggle website.The Dataset holds the records of 5731 Emails.The dataset has 2 columns representing the email content including subject,body and the other consists of two class 1(one) and 0(zero) , which represents Spam and non-spam or Ham Emails respectively.The Dataset is collected based on content-based filtering. where By Analyzing the email's content, such as its keywords, patterns and phrases to see if it matches any known spam patterns or features,

II. Data Cleaning:
The Raw dataset that is collected is taken for cleaning the data before it is used for analysis. Here.The Data here is Preprocessed ,Taking the stop words out. Stop words are the ones that appear a great deal in any text. For instance, the terms "the," "a," "an," "is," "to," etc. We learn nothing about the text's content from these words. Therefore, if we leave these words out of the text, it shouldn't matter. Duringthe pre-process, the unique characters and symbols are eliminated. White spaces and punctuation signs are employed to separate the text; these tools are referred to as tokenizers, and our dataset makes use of one. Tokenizers are used for extractracting the features, which are then input to the training model.

III. Data Analysis : After the Data cleaning the data is set for the analysis.The analysis of data is done using the Preprocessed dataset . The dataset , divided into Testing and training dataset . The Machine Learning Algorithms are tested on the obtained Dataset . The Various MLRegression Algorithms includes KNN , Ada Boost , Decision tree ,Random Forest etc .
For each of the Algorithms the Accuracy , RMSE ,and Confusion matrix are drawn using heatmaps and the comparative study has been performed.

IV. Data Exploration : The Result obtained from the analysis of Dataset using the Regression Algorithms are represented visually using the graphs , plots like bar ,chat line graph etc.

IV. RESULTS AND DISCUSSION

The following are the ML Classification Algorithms that are implemented :

1. Logistic Regression : An illustration of supervised learning is logistic regression. It is used to determine or forecast the likelihood that a binary (yes/no) event will occur.
The following result is obtained after using Linear Regression Algorithm on the dataset

Parameter	Values Obtained
MSE	0.06
MAE	0.06
R-Squared	0.91
RMSE	0.24
Accuracy	99.00%

Table 1 : Parameters of Logistic Regression

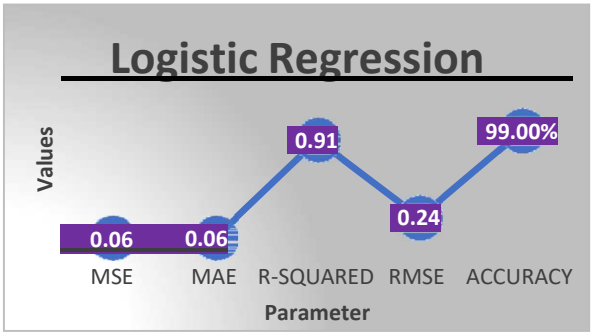

Figure 2 : Visual representation of Table 1

2. AdaBoost : An ensemble method in machine learning that uses the Boosting approach is called adaptive boosting. The weights are redistributed to each instance, with higher weights being given to instances that were mistakenly identified

Parameter	Values Obtained
MSE	0.04
MAE	0.04
R-Squared	0.88
RMSE	0.19
Accuracy	97%

Table 2 : Parameters of AdaBoost

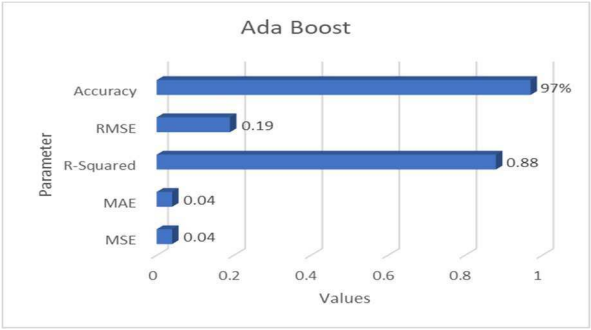

Figure 3: Visual representation of Table 2

3. Gradient Boosting : A machine learning method called GB is used, for classification and regression tasks. It is

an ensemble method that combines multiple weak predictive models, typically decision trees, to create a strong predictive model.

The following result is obtained after using Linear Regression Algorithm on the dataset

Parameter	Values Obtained
MSE	0.08
MAE	0.08
R-Squared	0.89
RMSE	0.28
Accuracy	97.00%

Table 3 : Parameters of Gradient Boosting

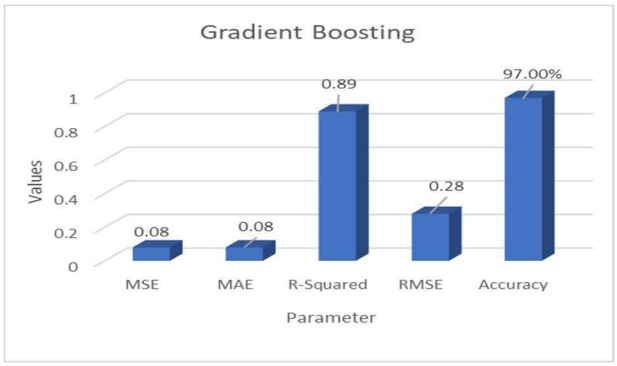

Figure 4: Visual representation of Table 3

4. Decision Tree : It creates a model in the form of a tree structure, where each internal node represents a feature or attribute, and each leaf noderepresents a class label or a prediction.

Parameter	Values Obtained
MSE	0.08
MAE	0.08
R-Squared	0.92
RMSE	0.29
Accuracy	98.00%

Table 4 : Parameters of Decision Tree

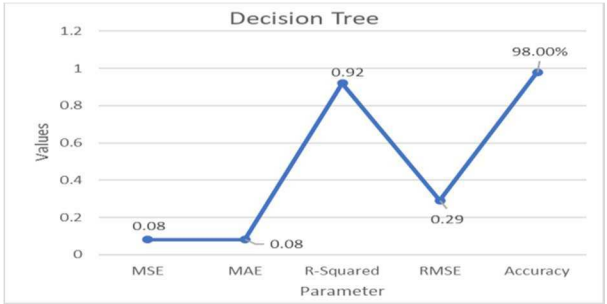

Figure 5 : Visual representation of Table 4

5. KNN : An approach to classification of Data that calculates a data point's likelihood of belonging to one group or the other based on which category the data points closest to it belongs to.

Parameter	Values Obtained
MSE	0.2
MAE	0.2
R-Squared	0.85
RMSE	0.45
Accuracy	91.00%

Table 5 : Parameters of KNN

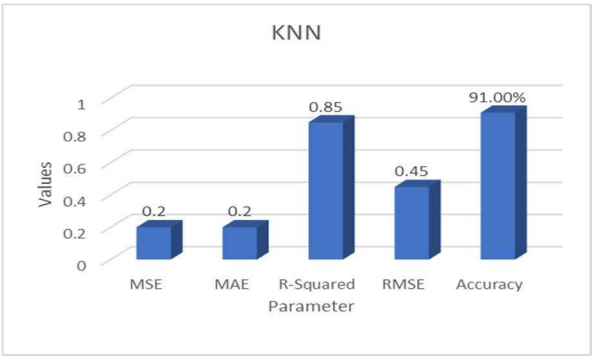

Figure 6 : Visual representation of Table 5

6. Random Forest : The random forest takes the prediction from each tree and bases its prediction of the final output on the majority votes of predictions , and improves the accuracy of the dataset.

Parameter	Values Obtained
MSE	0.05
MAE	0.05
R-Squared	0.92
RMSE	0.22
Accuracy	98.00%

Table 6 : Parameters of Random Forest

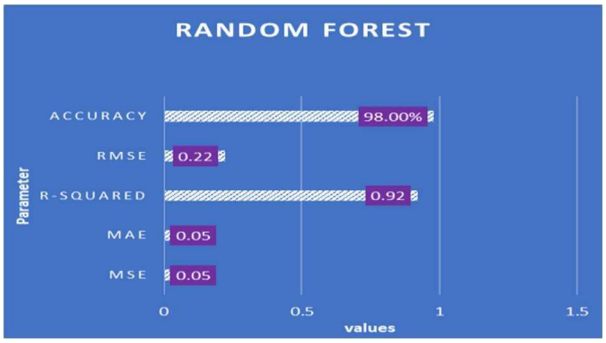

Figure 7 : Visual representation of Table 6

7. Naive Bayes : The Naive Bayes algorithm is a supervised method for the classification issues that is based on the Bayes theorem and is mostly employed in text categorization with a large training set.

Parameter	Values Obtained
MSE	0.04
MAE	0.04
R-Squared	0.92
RMSE	0.2
Accuracy	99.00%

Table 7 : Parameters of Naïve Bayes

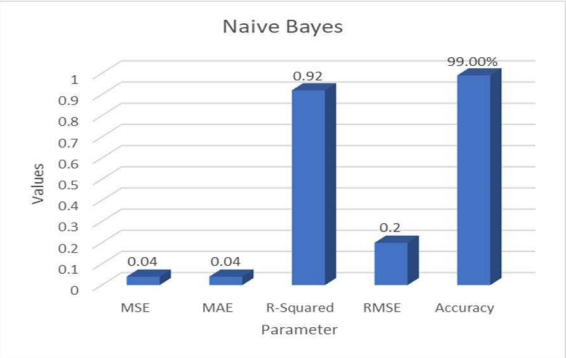

Figure 8 : Visual representation of Table 7

8. Voting Classifier : A voting classifier is a type of machine learning estimator that develops a number of base models or estimators and makes predictions based on averaging their results.

Parameter	Values obtained
MSE	0.02
MAE	0.02
R-Squared	0.89
RMSE	0.14
Accuracy	0.98

Table 8 : Parameters of Voting Classifier

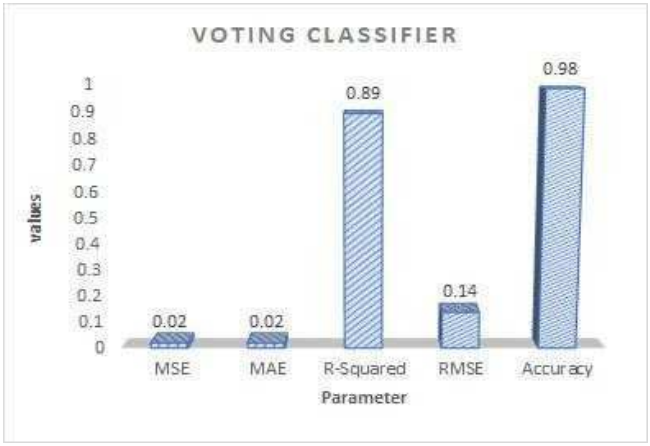

Figure 9 : Visual representation of Table 8

Algorithms	Accuracy (%)	RMSE
Naïve Bayes	99.8	0.2
Logistic Regression	99.5	0.24
Voting classifier	98.6	0.14
Random Forest	98.4	0.22
Decision Tree	98.2	0.29
Ada Boost	97.7	0.19
Gradient Boosting	97.3	0.28
KNN	91	0.45

Table 9 : Comparison table for Accuracy and RMSE of different ML Algorithms

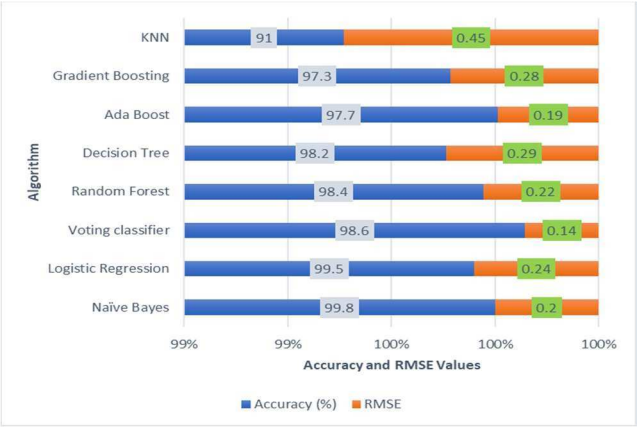

Figure 10 : Graphical representation of Table 9

The Table 9 , gives the clear picture of the Obtained Accuracy and RMSE of the ML algorithms respectively . Metrics like accuracy and RMSE are used to assess how well a predictive model is working.

The percentage of values that were accurately predicted out of all possible outcomes is known as accuracy. In this instance, a 99.8% accuracy indicates that 99.8% of the values were properly predicted by the model.

The average discrepancy between the values that were predicted and the actual values is measured by RMSE. The predictions of the model are more accurate in case of smaller the RMSE value is. An RMSE of 0.2 indicates that the model's predictions are typically 0.2 units off from the actual values.

From the Analysis , here we can see the Highest Accuracy of 99.8% for Naive Bayes Algorithm with the RMSE of 0.2 in turn is the best for the obtained dataset .

The confusion matrix is obtained and is represented in the form of Visualization techniques that includes Heatmap .

Hence the model best fits the Dataset correctly with the maximum value , Naïve Bayes stands top and is followed by the Logistic Regression and Voting classifier algorithm with the accuracy of 99.5% and 98.6% respectively.

V. CONCLUSION

The Machine Learning Algorithms such as Naive Bayes, Random Forest , Ada Boost etc., are applied for the dataset with the Content Based Filtering obtained from the Kaggle website , we conclude that naive bayes algorithm suits best with maximum accuracy of 99.8% and RMSE of 0.2.

The KNN algorithm is found to have least accuracy of 91% and RMSE of 0.45 , Indicating the not so best fit for the chosen Dataset pattern.

This approach of Email spam detection can be employed for the dataset with the Content Based Filtering , blacklisting and Whitelisting methods .

REFERENCES

[1] Spam Detection System Using Supervised ML , Abhila B1,Delphin periyanayagi ,Koushika ,Mabel Nirmala Joseph ,Dhanalakshmi , 021 International Conference on System, Computation, Automation and Networking (ICSCAN) | 978-1-6654-3986-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSCAN53069.2021.9526421

[2] An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection , Tasnia Toma, Samia Hassan, Mohammad Arifuzzaman , 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 8-9 July 2021

[3] SMS spam detection and comparison of various machine learning algorithms , Shubhangi Suryawanshi , Anurag Goswami , Pramod Patil , 978-1-7281-4392-7/19/$31.00 @ 2019 IEEE

[4] Email Spam Detection : An Empirical Comparative Study of Different ML and Ensemble Classifiers , Shubhangi Suryawanshi, Anurag Goswami, Pramod Patil , 9th International Conference on Advanced Computing (IACC) 978-1-7281-4392-7/19/$31.00 c 2019 IEEES.

[5] A Comprehensive Survey for Intelligent Spam Email Detectio , Asif karim , sami azam , bharanidharan shanmugam , krishnan kannoorpatti , and mamoun alazab , IEEEACCESS, date of current version December 4, 2019. DOI: 10.1109/ACCESS.2019.295479

[6] Detecting Spam in Emails , Applying NLP and Deep Learning for Spam Detection by Ramya Vidiyala Towards Data Science, Blog

[7] A Multivocal Literature Review on Growing Social Engineering Based Cyber-Attacks/Threats During the COVID-19 Pandemic: Challenges and Prospective Solutions , Mohammad Hijji , Gulzar Alam Year 2020

[8] 'What Is Machine Learning? 3 things you need to know', Accessed on: March 2021. [Online]. Available: https://www.mathworks.com/discovery/machine-learning.html

[9] Z. Ghahramani, 'Probabilistic machine learning and artificial intelligence,' Nature, vol. 521, no. 7553, 2015, pp. 452–459

[10] 'Tackling the poor assumptions of naive bayes text classifiers,' J.D.M. Rennie, L. Shih, J. Teevan and D.R. Karger, In Proceedings of the 20th international conference on machine learning (ICML-03),2003, (pp. 616- 623).