IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, April 2024

Email Spam Classifier

Chitra Malode[1], Pallavi Lanjewar[2], Priyanka Gomase[3]
MCA Department[1,2]
CSE Department[3]
K.D.K College of Engineering, Nagpur, Maharashtra, India
chitramalode.mca23@kdkce.edu.in[1], pallavilanjewar.mca23@kdkce.edu.in[2], priyankagomase@gmail.com[3]

Abstract: *Communication plays a major part in everything be it proficient or individual. Because of its widespread use, accessibility, affordability, and free services, email is a popular communication tool. The rise in email-based attacks is a direct result of email protocol weaknesses as well as the growing volume of electronic commerce and financial activities. One of the main issues with today's Internet is email spam, which can financially harm businesses and bother individual consumers. On the internet, spam emails are the main problem. Spammers find it simple to send emails that are filled with spam. Our inbox is flooded with several pointless emails from spam. We receive an overwhelming volume of spam emails every day, making it difficult and time-consuming for us to distinguish between them. Spam remains a problem despite all the efforts made to eradicate it. Furthermore, even valid emails will be removed from consideration when countermeasures become excessively sensitive. Filtering is one of the key strategies among the methods created to prevent spam. This research aims to explore machine learning algorithms and their application to our data sets. The optimal algorithm for email spam detection is chosen based on its optimal precision and accuracy.*

Keywords: Machine Learning, Spam Classification, NLP

I. INTRODUCTION

Even if social communication has evolved, mail remains one of the most widely used forms of communication. It may be used in both formal and informal contexts, and these days, everyone uses emails to communicate official information. The growth of invention has also led to an expansion of cybercrime. Email is preferred over other physical tactics because to its many benefits, including minimal expenses, security of the information being transferred, and irrelevant time delay during transmission. Nevertheless, a few problems impede the efficient use of emails. Among them is spam mail. Spam can be sent from anywhere on the earth by clients with false eagerness that has gained access to the Internet. Spam refers to unsolicited and impulsive emails that are sent to recipients who are not in need of them. These spam emails are sent in bulk to a large number of recipients and contain bogus content that typically links to phishing attacks and other threats. In order to tackle this problem, machine learning techniques have evolved into effective tools for detecting spam in emails.

The mail spam classifier's main goal is to accurately identify and categorize incoming emails as spam or legitimate (ham). The constantly evolving nature of spam has limited the adequacy of traditional rule-based techniques. Machine learning provides a more dynamic and adaptable method by utilizing patterns and highlights extracted from large-scale email databases. A machine that can identify and categorize spam. We use data-mining classification calculations to try and separate designs so that we can label the emails as SPAM or HAM.

II. LITERATURE SURVEY

[I] Zeeshan Bin Siddique,Mudassar Ali Khan ,IkramUd Din ,Ahmad Almogren ,Irfan Mohiuddin,ShahNazir (2021) Published Research Article on Machine Learning-Based Detection of Spam Emails at Hindawi Scientific Programming
[I] They described a focused Introduction of the model to be implemented in the classifier
[II] Hazel Murphy presented final report on Email Spam Filter using Machine Learning (2021) at Institute of Techonology Carlow
[I] Here, the brief description about the methods to be used is given in the Introduction

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, April 2024

[III] Sanket Sonowal, Nikhil Kumar, Nishant has conducted research on " Email Spam Detection using Machine Learning Algorithms" at 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) [3]

They have well-explained the methodology to be used to classify and differentiate the emails that are spam or not. Their study focuses on email spam filtering and provides insights into feature selection and use of different algorithms applicable to email spam detection using Machine Learning.

III. PROPOSED METHODOLOGY

The issue resolved in this extend is the expanding sum of spam emails that are attacking client inboxes without their assent, using more network capacity, and causing budgetary harm to companies. In spite of measures taken to kill spam, we proposed in the Machine Learning Models such as Naïve Bayes, SVM, KNN Models that have the most elevated accuracy when compared to the existing framework. Following is the flow in which existing system is to be implemented:

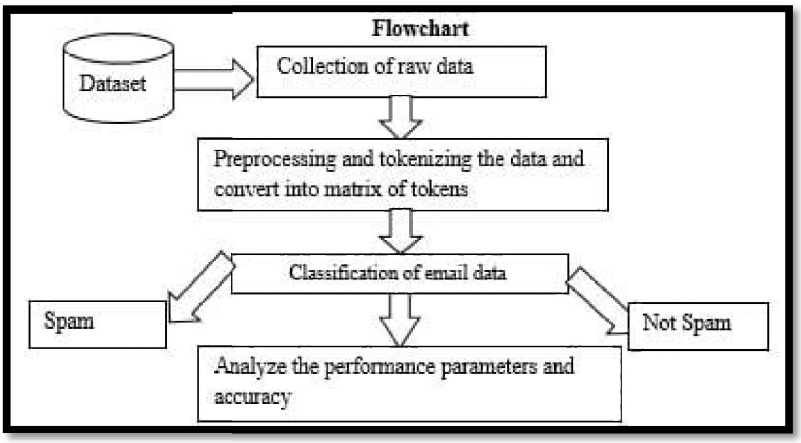

Fig 1: Flowchart for Email Spam Detection

A. Data Collection: Data plays an crucial role when it comes to classification, the more the information the more the precision will be. The data utilized in this extend is totally open-source and has been taken from various resources like UCI.

B. Data Preprocessing: The messages have to be pre-processed for removing the extra or unwanted data consisting of undesirable puncuation, accentuation, stop words etc

1. Overall data processing: It comprises of two fundamental tasks

a. Dataset Cleaning: It incorporates different tasks such as removing outliners, null value removal, invalid esteem evacuation, removal of unwanted highlights from data.

b. After cleaning the data, the datasets are merged to frame a single dataset containing two features(text, label). Data cleaning, Data merging these methods are totally done using Pandas library.

2. Textual Data Processing: The overall data is processed only to datasets, the textual processing is done to both user input data, training and testing data.

a. Tag removal: Expelling all sorts of tags and unknown or undefined characters from content utilizing regular expressions through Regex library.

b. Sentencing, tokenization: Breaking down the text(email/SMS) into sentences and at that point into tokens(words).

c. Stop Word Removal: While using the NLP i.e. Natural Language Processing, our main objective is to perform analysis so that a computer can respond to text properly. Stop words such as of , dg ,sk , … are removed using stop words NLTK library of python.

d. Lemmatization: Words are changed over into their base forms using pos-tagging and lemmatization. This process gives key-words through entity extraction. This process is done using chunking in Regex and NLTK lemmatization.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, April 2024

e. Sentence formation: The lemmatized tokens are combined to shape a sentence. This sentence is basically a sentence converted over into its base form and evacuating stop words. At that point all the sentences are combined to form a text.

3. Bag of Words: "Bag of Words (BOW) is a method of extracting features from content records. Later, these features can be used for training Machine Learning Algorithms. Bag of Words makes vocabulary of all the one kind of words display in all the report in the Training dataset."It is used mainly in content or text classification. A bag of words represents content of text in a numerical form. The two things required for Bag of Words are

i. A vocabulary of words known to us. .

ii. A way to measure the presence of words.

4. Term Frequency-Inverse document frequency : Term Frequency-inverse document frequency of a word is a estimation of the significance of a word. It compares the repentance of words to the collection of reports, documents and calculates the score. The TF-IDF prepare comprises of different activities listed below.

i) Term Frequency: The count of appearance of a particular word in a document is called term frequency

$$\boldsymbol{tf(t, d) = count\ of\ t\ ind/\ number\ of\ words\ in\ d}$$

ii) Document Frequency: Document frequency is the count of documents the word was detected in. We consider one instance of a word and it doesn't matter if the word is present multiple times.

$$\boldsymbol{df(t) = occurrence\ of\ t\ in\ documents}$$

iii) Inverse Document Frequency: It includes Inverse Document Frequency which reduces the importance of words/terms that has high recurrence and increases the importance of words/terms that are rare.

$$\boldsymbol{idf(t) = N/df}$$

Finally, the TF-IDF can be calculated by combining the term recurrence and inverse document frequency.

$$\boldsymbol{tf\_idf(t, d) = tf(t, d) * \log\ (N/(df + 1))}$$

C. CLASSIC CLASSIFIERS: Classification is a form of data analysis that extracts the models describing vital information classes. A classifier or a model is developed for prediction of class labels for illustration: "A loan application as risky or safe." Data classification is a two-step - learning step (construction of classification model.) and - a classification step

1. NAIVE BAYES: The Naïve Bayes classifier is quick classification used for supervised learning. The Naive Bayes is based on theBayes Hypothesis also known as Bayes Theorem and works on the dependent events which have probability of the event which is going to happen in the future that can be identified from the same event which occurred previously. Bayes theorem assumes that highlights are independent each other. Naïve Bayes classifier technique can be used for classifying spam emails as word probability plays primary role here. If there is any word which occurs frequently in spam but not in ham, at that point that email is spam. Every time the Naive Bayes calculates the probability of each class and the class having the maximum probability is then chosen as an output. Naïve Bayes always give an accurate result.

2. Support Vector Machine: The Support Vector Machine (SVM) is a popular Learning algorithm,.the Support Vector model is used for both classification problems in Machine Learning techniques and regression challenges. The Support Vector Machines totally founded on the idea of Decision points. In the SVM algorithm, we mark each data value as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. The Main resolution of Support Vector Machine algorithm is to create the line or decision boundary.

3. Decision Tree: Decision Tree is a graphical representation ***for getting all the possible solutions to a problem/decision based on given conditions.*** To improve the performance, the decision tree algorithms is used in email classification. A decision tree is a flow chart like construction, where it starts with the root node, which expands on further branches and constructs a tree-like structure. A decision tree splits into sub-trees based on answers (Yes/No) by simply asking a question.

4. K- NEAREST NEIGBOUR: K-nearest neighbour is non-parametric algorithms that identifies the nearest neighbours like distinguishing the spam mails. This algorithm has some data point and data vector that are isolated into few classes to foresee the classification of new sample point. K-NN algorithm stores all the accessible information and classifies a unused data point based on the similarity. This means when new data shows up at that point it can be easily classified into a well suite category by using K- NN algorithm.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, April 2024

D. ENSEMBLE LEARNING METHODS: Ensemble methods in machine learning is a method that builds a collection of decision tree and combines them to get final output. The main aim of ensembling learning method is to improve the accuracy of result by using combination of multiple models

1. RANDOM FOREST CLASSIFIER: Random forest classifier is machine learning ensemble tree classifier which is a process of *combining multiple classifiers to solve a complex problem and to improve the execution of the model. It* comprises of distinctive sorts of choice of trees that are of diverse shapes and sizes. The arbitrary examining of the training data when building a tree. Instead of depending on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output

2. BAGGING : Bagging fits base classifiers each on random sub sets of the original data sets and then combined their individual calculations by voting or by averaging) to form a final prediction. "Bagging is a mixture of bootstrapping and aggregating. Bagging= Bootstrap aggregating.

3. Bootstrapping: Bootstrapping helps to lessening the variance of the classifier and it also decline the overfitting by just resampling the data from the training data with same cardinality as in original data set. High variance is not good for the model. Bagging is very effective method for limited data, and by just using samples you are able to get estimate by aggregating the scores .

IV. CONCLUSION

More than 300.4 billion emails are exchanged daily, about 57% of these are just spam emails. With the increase usage of emails, this study focuses on using automated ways to determine spam emails. Our study uses various machine learning and deep learning algorithms to determine them. By using the power of supervised learning algorithms such as Naive Bayes, Support Vector Machines, and KNN, and by preprocessing the text data using techniques such as tokenization, stop- word removal, and stemming, it's possible to build accurate and reliable spam filters that can automatically determine and filter out unwanted emails. These techniques can also be extended to handle more complex spamming strategies similar as phishing attacks and shaft phishing. Overall, in the proposed models Naïve Bayes having the delicacy of 95.24% SVM having 96.90% and KNN having 96.20%.

V. ACKNOWLEDGMENT

A substantial activity can only be successfully and satisfactorily completed with the involvement of diverse personal effort from all angles, both explicit and implicit. Wide-ranging, worthwhile reading activities result in significant knowledge gains from books and other informational sources, but real competence comes through related learning tasks and experience. We ardently with extent, both modestly our heartfelt gratitude to all those who provided timely and honest assistance in making this project a success.

We sincerely thank and express our gratitude to our project guide, Prof. PRIYANKA GOMASE for their expert guidance in achieving the objectives mentioned above. We express our gratitude to respected Dr. ANUP BHANGE, Head of Department of Master Of Computer Application (MCA) and other staff members for guiding us and giving their valuable suggestions.

REFERENCES

[1] Mrs. Anitha Reddy, KanthalaHarivardhan Reddy , A. Abhishek , Myana Manish , G. Viswa Sai Dattu , Noor Mohammad Ansari conducted a research on Email Spam Detection Using Machine Learning at Journal of Survey in Fisheries Sciences(2023)

[2] DarshanaChaudhari, DeveshriKolambe, RajashriPatil, Sachin PuranikU.G. Students, Department of Information Technology, SSBT's College of Engineering and Technology, Bambhori, Jalgaon, India, conducted research on EMAIL SPAM DETECTION USING MACHINE LEARNING AND PYTHON published in proceedings International Journal of Research Publication and Reviews (2022)

[3] Zeeshan Bin Siddique,Mudassar Ali Khan ,IkramUd Din ,Ahmad Almogren ,Irfan Mohiuddin,ShahNazir (2021) Published Research Article on Machine Learning-Based Detection of Spam Emails at Hindawi Scientific Programming

[4] Hazel Murphy presented final report on Email Spam Filter using Machine Learning (2021) at Institute of Techonology Carlow

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, April 2024

[5] SanketSonowal,Nikhil Kumar,Nishanthas conducted research on " Email Spam Detection using Machine Learning Algorithms" at 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)
[6] VATHUMALLI SRI GANESH, VATTIKUTI MANIDEEP SITARAM conducted research on "Spam Detection using Machine Learning and Natural Language Processing"