IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

IJARSCT
Impact Factor: 6.252

Volume 2, Issue 4, May 2022

Email Spam Filtering using Machine Learning

Sagar Kaulagi[1], Satej Kumbhar[2], Kartik Gambhire[3], Abhijeet Sarvade[4], Prof. Umesh Nanavare[5]
Students[1,2,3,4] and Professor[5]
Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Abstract: *In todays world, all activities depend upon the internet. In that Receiving Spam email send messages is a major problem. Many times, this kind of mail contains viruses and hacking links and they affect our system. For solving this kind of problem, we need some method that can filter spam mails and non spam emails. In this paper, we presented one machine learning method that filters spam and non-spam emails. Our algorithm generates the dictionary and features vector and trains them with a machine learning algorithm. Email is one such communication medium that comes to mind when we think of secure communication. As the popularity of email increases, the number of unsolicited data has also increased rapidly. A lot of unwanted stacks of emails called as Spam has created a need for further development Nowadays Machine learning methods have been able to detect and filter out spam emails. The purpose of this current project is to receive a spam email in the morning or effectively using the Multinomial Naïve Bayes method. Naïve approach is a machine-readable algorithm used to classify sample email as spam or not. This filter can be used by other email service providers as fully functional spam filters.*

Keywords: Unsolicited Data, Spam Emails, Machine Learning, Multinomial Naïve Bayes Route, Supervised Learning.

I. Introduction

Today, a large number of unwanted emails called spam have created a major problem for the Internet. A person who sends spam mail is referred to as spam. Spam senders collect emails on websites, chat rooms and various viruses. Spam email reduces storage capacity and network bandwidth. It greatly affects email servers, network bandwidth, CPU and user time. The presence of spam email increases every year and is responsible for more than 77% of global email traffic. It has been found that many users have experienced financial losses that have been the victims of email scams and other unscrupulous processes sent to spam email senders who pretend to be from reputable companies for the purpose of knowing sensitive personal information such as One Time. Passwords (OTPs), debit / credit card numbers. Nowadays, email provides many ways to send millions of advertisements at no cost to the sender. As a result, many unsolicited bulk e-mail, also known as spam e-mail spread widely and become serious threat to not only the Internet but also to society. For example, when a user receives a large amount of email spam, the chance of the user for getting to read a non spam message increases. As a result, many e-mail readers have to spend their time removing unwanted messages. E-mail spam also may cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content. Over the past many years, many approaches have been provided to block email Spam. 1In the proposed system, we are working on the above issues like how we can reduce the problems which are created by the email spammers and protect the users' sensitive data from these fraudulent mails.We are introducing a machine learning model which uses the Naïve Bayes classifier algorithm which classifies the email coming in to spam and non spam mails. Due to this Many of the issues related to the spam mails can be reduced as well as the sensitive information of the email users can be protected. Naïve Bayes algorithm is a group 5of classification algorithms. Naïve Bayes is not a single algorithm but a family of the algorithm where all of them share common principles i.e. every pair of features being classified is independent of each other.

Many of the Machine Learning algorithms used in history for email spam filtering. The naïve bayes is easy to implement and has fast evaluation to filter email spam. There search objectives are:

1. Implementation of Naïve Bayes Algorithm on a dataset
2. Evaluation of performance of Naïve Bayes algorithm.

DOI: 10.48175/IJARSCT-3999

II. LITERATURE SURVEY

In this paper the authors used the Naive Bayes algorithm to filter email spam into two databases. For database performance measurement use data accuracy, memory and F scale. The Naive Bayes algorithm is very popular with commercial and open source filters. This is due to the simplicity of the data filtering algorithm. Only a short training period is required to filter email spam. Training data is provided in the form of raw emails and classifies emails as spam and non-spam. Today spam senders have the power to launch spam campaigns. These spam can be widely distributed using malware and botnets by spam. When a user receives and opens a spam email, sensitive user information or data could be compromised by botnet senders via viruses and viruses.

Several machine learning algorithms have become a growing problem over the years. About 70% of all emails are spam. Like web extensions, the problem of email spam is also growing. According to [1], an average of 10 days a year was found to be at risk for spam analysis. Spam is an expensive problem that can be very costly in the coming years to reduce bandwidth providers. Spam is an important issue that attacks email presence. Therefore, it is very important to classify unwanted email into various recommended ways to identify and classify email messages as spam or non-spam or email, as well as to determine the effectiveness of the algorithm. The speed of machine learning is very high. Many algorithms are processed to classify unsolicited emails that are widely used and analyzed among them as vector machines. Naive Bayes Tree Determining the division of emotional networks is divided into categories. In this article we have tried our algorithms: Naive Bayes, Bayes Net, Vector Machine Support (SVM), Tree Work (FT), J48, Random Forest and Random Tree.

In this paper, the author's intention is to understand the data features that affect the performance of the non-critical beams. This approach uses the simulation of Monte Carlo that allows for accurate reading of sections in randomly generated problems of several classes. "Analyze the impact of propagation entropy on the segmentation error, which shows that the propagation element of low entropy reflects the good performance of the Naive Bayes". Another surprising result is that the accuracy of the Naive Bayes is not directly related to the level of dependence of the factor measured as phase- conditional information between the elements.

III. SYSTEM ARCHITECTURE

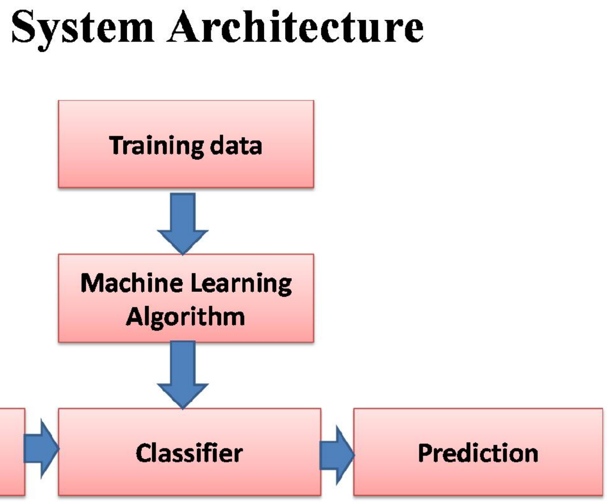
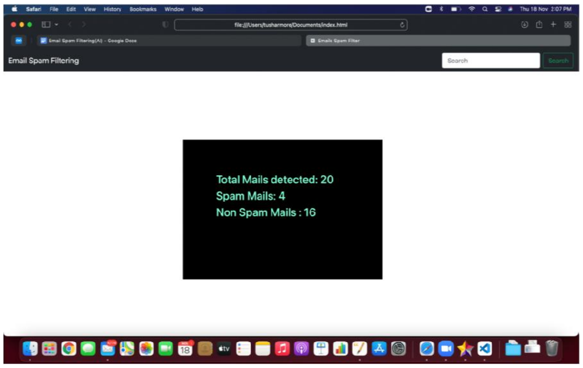


Figure 1: System Architecture

First data will be loaded as for training the model, using naïve byes algorithm model will be built. Once model built then email as given as a input for the testing if it is spam them that will be detect as spam, if not then not spam.

IV. RESULTS

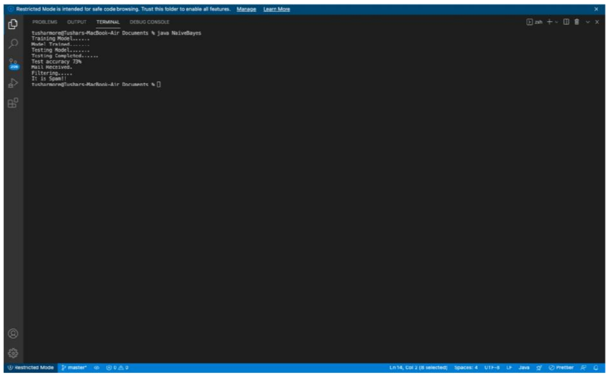
We have developed the machine learning model in python. This model uses the same Naïve Bayes Classifiers approach to classify the spam mails. It displays the count of the spam and non spam mails. We used the google API to connect our account to this model. Some results/outputs of the model is attached below:


Safari File Edit View History Bookmarks Window Help
Thu 18 Nov 2:07 PM
file:///Users/tusharmore/Documents/index.html
Email Spam Filtering(AI) - Google Docs
Emails Spam Filter
Search
Search

Email Spam Filtering

Total Mails detected: 20
Spam Mails: 4
Non Spam Mails : 16

Figure 5: Web Based Client for Visualization


Restricted Mode is intended for safe code browsing. Trust this folder to enable all features. Manage Learn More
PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
tusharmore@Tushars-MacBook-Air Documents % java NaiveBayes
Training Model......
Model Trained.......
Testing Model.......
Testing Completed......
Test accuracy 73%
Mail Received.
Filtering.....
It is Spam!!
tusharmore@Tushars-MacBook-Air Documents % 
```

Figure 6: Output

V. Future Scope

Though the algorithm is very famous because of its simplicity and effectiveness but it has also some up and down sides below are some advantages and disadvantages of Naïve Bayes Algorithm are given: Advantages:

- It is simple and easy to implement.
- It doesn't require as much training data.
- It handles both continuous and discrete data.
- It is highly scalable with the number of predictors and data points.
- It is fast and can be used to make real-time predictions.
- It is not sensitive to irrelevant features.

VI. Disadvantages

- The Naïve Bayes ignore the meaning of sentences. It works on each word. This limits the effectiveness of the algorithm in real-time use cases.
- Provides zero chance of class division of its category in the test database was not available in the training database. This problem can be solved continuously using the laplace smoothing method.
- This algorithm works well in the theory section, but in real time it does not give the full value of the correct output.

VII. Conclusion

From recent days the spam emails are increasing day by day and creating a problem for the user so with a spam detector, we will determine which spam emails or not, with this efficiency users will increase. We use the Naïve Bayes section which will provide a possible indication of this and will determine whether or not the mail is spam is based on the results shown.

VIII. Acknowledgement

It is a matter of great pleasure for us to submit this paper on "Email spam filtering using machine learning as a part of curriculum for award of "Bachelor's in Engineering (Computer)" degree of Savitribai Phule Pune University. Firstly, we would like to express my gratitude to my guide **Prof. Umesh Nanavare,** for his inspiration, guidance,constant supervision, direction and discussion in successful completion of dissertation. We are also thankful to him for his support in providing their port format, making flexibletime schedules for assignments, test, and lectures to all students during entire year. We are grateful to Head of Department **Dr**. **R. H. Borhade**, for his valuable support and guidance.

We are thankful to my **Principal Dr. A. V. Deshpande** and to all our staff members who encouraged us to do this, we also extend our thanks to all our colleagues those who have helped us directly or indirectly in completion of this project.

References

[1]. M. Awad, M. Foqaha, Email spam classification using hybrid approach of RBF neural network and particle swarm optimization,Int. J. Netw.Secure.Appl.8 (2016).

[2]. D. M. Fonseca, O. H. Fazzion, E. Cunha, I. Las-Casas, P. D. Guedes, W.Meira, M. Chaves, Measuring characterizing, and avoiding spam traffic costs, IEEE Int. Comp.99(2016).

[3]. Visited on jan 15 ,2022, Kaspersky Lab Spam Report,2017,2012.
https://wwwsecurelist.com/en/analysis/204792230/Spam_Report_April12.

[4]. Megha Tope, Email Spam Detection using Naïve Bayes Classifier(2017).

[5]. Rish, An empirical study of the nssaïve Bayes classifier, T. J. Watson Research Center.

DOI: 10.48175/IJARSCT-3999