PROJECT SYNOPSIS

ON

A Comparative Analysis of Machine Learning Algorithms for Spam Email Prediction

A synopsis submitted for partial fulfilment of the requirement for the degree of

Masters in Computer Application

Specialized in Cloud Computing and Information Security

Under

Assam Down Town University, Guwahati

Submitted By

Name:Raju Das	Name: Raja Alivardi Alam	Name: Hirakjyoti Lahkar
Roll No:ADTU/2018-23/ICA/020	Roll No:ADTU/2018-23/ICA/030	Roll No:ADTU/2018-23/ICA/005
Email: rajud878@gmail.com	Email: raja21525aa@gmail.com	Email: hlahkar94@gmail.com
Phone: 8837412882	Phone: 7002906681	Phone: 6001463216

Int.MCA (CTIS) 10th Semester

Under the Guidance of

Dr.Banani Das

Assistant Professor

Faculty of Engineering & Technology

Contents **Page No**

Introduction

Email is widely utilised in today's digital world. Email has established itself as a popular method of communication be it business or healthcare. There are two categories of email: HAM and SPAM. Emails that are used to hurt consumers by stealing their important information, wasting their time, and draining their resources are known as spam emails or trash mails. HAM emails, on the other hand, are a type of email that is not spam. Throughout the decade, the amount of spam emails has increased, and many email service providers have created a number of strategies to stop them. Mail filtering is one of the most crucial and well-known strategies among all others. For this, a variety of machine learning and deep learning approaches, including Naive Bayes, Random Forest and Decision trees have been used. This paper surveys the machine laerning techniques that are used for spam filtering. Also a comprehensive comparison of these techniques will be made based on their precision and accuracy.

Literature Survey

The paper [1] talks about different ways to find and stop spam emails. The best method is called Multinomial Naïve Bayes, but it's not perfect because it assumes that each word in an email is not related to any other words, which isn't always true. So sometimes it may make mistakes and not catch all the spam.

The paper also talks about using a bunch of different methods together to do a better job at stopping spam. But the project in the paper only tested a small number of emails, so it might not work as well on a lot of emails.

The author suggests that one way to do a better job at stopping spam is to only allow emails from trusted sources to get through. They also say it's important to be really good at telling which emails are spam and which aren't. Big companies can use this method to stop spam from getting to their employees.

Overall, the paper talks about how hard it is to stop spam emails, but gives some ideas for how to do a better job.

Feasibility Study

A feasibility study, usually referred to as a feasibility analysis, is a method of determining whether or not a project plan will be successful. A detailed feasibility study evaluates the project's viability in order to determine whether it can be moved forward or not. The following possible study can be completed as part of the suggested project:

Operational feasibility: The project requires a team with the necessary skills and knowledge to implement and compare machine learning algorithms. The team will also require access to email datasets for training and testing the algorithms. These resources are readily available, and the project is operationally feasible.

Technical feasibility: The project requires knowledge of machine learning algorithms and programming skills to implement and compare the algorithms. The required hardware and software resources, such as a computer with a suitable configuration and programming tools, are readily available and affordable, making the project technically feasible.

Economic feasibility: The cost of hardware and software resources required for the project is minimal also the dataset for training and testing the algorithms is publicly available and free of cost. Therefore, the project is economically feasible.

Objectives

The objective of the proposed project can be summed up from the given points:

1. To implement different classifying for spam email detection.

2. To preprocess and prepare the dataset for training and testing the algorithms.

3.To evaluate the performance of the implemented machine algorithms using metrics such as accuracy, precision.

4.To compare the performance of the different machine learning algorithms for spam email detection.

5.To identify the most effective machine learning algorithm for spam email detection based on the evaluation metrics.

6.To contribute to the development of a more robust spam email filter system that can help protect individuals and organizations from potential harm.

Problem statement

Spam emails are used to hurt consumers by stealing valuable information and waste their time. According to Statista , the estimated daily spam volume was 88.88 billion spam emails per day in September 2021. Given, the scenario, traditional rule based methods for spam email detection are becoming less effective in detecting new forms of spam emails. Therefore, there is a need of more advance techniques, such as machine learning and deep learning algorithms to detect and filter out spam emails. However with the avaibility of different machine learning algorithms, it is important to identify the most effective algorithm for spam email detection. The aim of this project is to implement and compare different machine learning algorithms and identify the most effective algorithm for accurate and effecient spam email detection.

Review

1. We have gathered information on the classification algorithms which we will be utilizing.

2. Information about spam email dataset.

3. Started working on the Logistic Regression Classifying Algorithm.

4. Software that is needed for implementing the project.

Methodology

We will use a publicly available data set of spam and non-spam emails to train and test the machine learning algorithms. The data set will be preprocessed to remove any irrelevant features and to convert them into format which is suited for the algorithms. We will then implement and compare different machine learning algorithms such as Logistic regression, Random Forest and Support Vector Machines (SVM). The algorithms will be evaluated based on metrics such as accuracy and precision.

Expected Outcome

The results will show the performance such as accuracy and precision of different machine learning algorithms for spam email detection. The results will be presented in the form of a comparison table and visualization to facilitate an easy understanding of the performance of different algorithms. We will identify the most effective algorithm for spam email detection based on the evaluation metrics.

Facilities required for proposed work

1. Python’s IDE: Annaconda
2. Google collab/ Kaggle
3. OS: Windows 10 or higher / Linux
4. Memory : Minimum 4 gb or above
5. Processor : atleast intel core i3 or higher
6. Stable Internet connection

Future Work

Although there are numerous machine learning and deep learning techniques for filtering spam email. These techniques might aid spammers or attackers in getting over the current filtering mechanism, which could be problematic. Therefore, in the future, new techniques may be developed, such as hybrid models or modifications to existing machine and deep learning models, such as "Improving the process of selecting the most important information from email and classifying them as spam or ham," which may be chosen by many organisations.

Conclusion

From a variety of machine learning algorithms that are currently available, we can draw the conclusion that the statistics of receiving spam emails can be decreased very effectively if the right machine learning algorithm is implemented, one that offers us higher accuracy rates and informed precision.

Bibliography

- https://www.googlescholar.com/
- https://www.researchgate.net/
- https://www.asana.com/
- https://mackeeper.com/
- https://statista.com/
- https://wikipedia.com/

Reference

[1]Email Spam Detection Using Machine Learning Algorithms by Nikhil Kumar,Sanket Sonowal and Nishant