Study of SPAM Email Detection

G.Mohan Sai Krishna[1#] K.Eswar Teja[2#] D.Harshavardhan Reddy[3#]

[1]*mohansaikrishnagundluru@gmail.com*
[2]*eswartejakotu@gmail.com*
[3]*harshavardhanreddy9640@gmail.com*
[#]*Department of Computer Science and Engineering*
Prathyusha Engineering College, Tiruvallur, Chennai- 602025

Abstract— A Many businesses and people now have easier ways to communicate thanks to electronic mail. Spammers who send unwanted emails use this technique to their advantage in order to make false gains. With machine learning algorithms that are enhanced using bio-inspired techniques, the goal of this paper is to show a method for spam email detection. A literature review is conducted to investigate the effective techniques used on various datasets to produce excellent results. On seven distinct email datasets, extensive research was conducted to apply machine learning models using Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and Multi-Layer Perceptron, along with feature extraction and pre-processing. To enhance the effectiveness of classifications, algorithms were put into place. Overall, the greatest performance was shown by logistic regression. Our findings are compared to those of other machine learning models in order to determine which model is the most appropriate.

Keywords— spam, email detection, machine learning.

I. INTRODUCTION

Major approaches adopted towards spam filtering include text analysis, white and black lists of domain names and community-based approaches. Text analysis of contents of mails is a widely used approach towards the spams. Many solutions deployable on server and client sides are available. Naive Bayes is one of the most popular algorithms used in these approaches. Spam Bayes and Mozilla Mail spam filter are examples of such solutions. But rejecting mails based on text analysis can be serious problem in case of false positives. Normally users and organizations would not want any genuine e-mails to be lost. Black list approach has been one of the earliest approaches tried for the filtering of spams. The strategy is to accept all the mails except the ones from the domain/e-mail ids. Explicitly blacklisted. With newer domains entering the category of spamming domains this strategy tends to not work so well. White list approach is the strategy of accepting the mails from the domains/addresses explicitly white listed and put others in a less priority queue, which is delivered only after sender respond a confirmation request

II. DATA

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

A. Text Data

This is the most important type of data used in spam detection systems. Text data includes the email's subject, body, and header information. Text data is used to train machine learning algorithms and to identify patterns or features that indicate whether an email is spam or not.

B. Metadata

This includes information about the email's sender, receiver, and timestamp. Metadata can be used to identify patterns or features that indicate whether an email is spam or not. For example, an email from a known spammer or a high number of recipients can be a sign of spam.

C. Behavioral data

This includes information about how users interact with emails, such as whether they open or delete them, or whether they mark them as spam. Behavioral data can be used to improve the performance of spam detection systems by providing feedback on the system's performance.

D. Network data

This includes information about the IP addresses and domains associated with the email, such as the email server and the email client. Network data can be used to identify patterns or features that indicate whether an email is spam or not.

E. External data

This includes information from external sources such as blacklists of known spammer IP addresses or domains, or information from other spam detection systems. External data can be used to improve the performance of spam detection systems by providing additional information about the email.

III. LITERATURE SURVEY

Spam prevention is often neglected, although some simple measures can dramatically reduce the amount of spam that reaches your mailbox. Before they are able to send you spam, spammers obviously first need to obtain your email address, which they can do through different routes.

Machine learning algorithms for email classification such as Neural Network (NN), Support Vector Machine (SVM), JDecision Tree based classifier, Naïve Bayes.In this paper work We proposed the basic three steps which are common in every classification process.

The first step is pre-processing in which the given text is converted into tokens and this step is also used for removal of stop words. The second step is learning process and, in this feature, set is built which is very much necessary for the classification of emails.

The last step is classification of email as ham or spam by using efficient algorithm. Algorithms like support vector machine, logistic regression, regression trees and random forest are considered for classification. They used the Phishing Corpus dataset and with the help of Bag of words as feature extraction approach classified the email as ham or spam. In his study, they did not mention the different tools for reduction methods for email classification.

Proposed a solution to increase the accuracy of Naive Bayes and reduce the false positive rate. Naive Bayes is a supervised machine learning algorithm which is based on Bayes theorem and can be used as a probabilistic model for classification of emails.

Although Naive Bayes classifier provides higher accuracy but still spammers are able to bypass the filter by using leetspeak and diacritics.

Leetspeak is a coded spelling system and language used in very informal communication on the internet, featuring letters combined with numbers or special characters in place of letters that they may resemble, and including inventive misspellings, jargon, and slang. Diacritic is a sign, such as an accent or cedilla, which when written above or below a letter indicates a difference in pronunciation from the same letter when unmarked

IV. CLASSIFICATION OF EMAILS

Classified emails based on their titles into four different categories: Sexual, financial, and marketing applications. They mainly focus on the features extracted only from the email header message. They also presented a novel

approach for filtering based on classified decision trees (DT), which involves implementation of the Decision Tree methodology to all the categories depending upon the attributes (features) that are collected from the header of e-mail. The retrieved characteristics from the sender's field are listed as the title of the mail, the date on which it was sent, and the size of the email.

Typically, spam contains advertisements for dubious products and services such as "Make Money Fast" schemes, multilevel marketing, illegally pirated software, and foreign bank scams. Spam may also contain offers to sell real estate, medicine, loans, and investments.

Spam, or unsolicited email, is widely regarded as a serious threat to the Internet, as it floods users' inboxes and costs businesses billions of dollars in wasted bandwidth. Spam's global productivity cost increased by $2 billion to $132 billion in 2010. Additionally, spam causes a number of negative consequences, including an overflow of email storage capacity, which disables the server's ability to receive new emails, as well as a slow response time from the server and insufficient system resources.

Email spam wastes the user's time in containing and deleting unwanted emails, results in the loss and/or delay of critical or emergency email messages and degrades Internet bandwidth and performance.

Additionally, spam is one of the most effective methods of spreading malicious programs, worms, and Trojan Horses that corrupt computers and operating systems. Additionally, spam contributes to an increase in the proportion of people exposed to fraud, resulting in the loss of millions of dollars worldwide each year.

V. Spam Filtering Techniques

The growing number of email users has irritated the public over the rise in spam. Email spam is inconvenient for users and costs email server owners and Internet service providers a lot of money to classify. There are numerous spam filtering techniques that fall into several categories. This study will summarize it.

User-defined filters: In this technique, filters automatically determine and remove spam messages based on user-defined rules. These include establishing guidelines for the acceptable subject matter and sources. For example, the user can configure his filter to reject all emails that contain a particular word in the header or emails from senders.

Header filters: This technique relies on inspecting the headers of incoming emails for forgery indicators. Numerous spammers

fabricate these headers in order to conceal their identities or locations. Normally, a header contains a wealth of information, including the recipient, subject fields, and sender; it also contains information about the servers that delivered this email, referred to as the relay chain. A good header filter can detect the forged header. Not all spammers, however, forge this information.

Language filters: This technique handles any email that is not written in the language selected by the user. Whether or not the user uses email to communicate with anyone in a foreign language, this technique is beneficial for non-native speakers.

Content filters: This technique is based on a set of rules that analyse the text of emails to determine whether they are spam or not. Unfortunately, this technique has some drawbacks, as these filters may exclude useful emails such as newsletters and other emails that the user specifically requests.

VI. Related Works

In this study, Unwanted email messages were first identified as a problem in a 1975 Internet Request for Comments. Once spam or unwanted email becomes prevalent, and the emergence of this problem and the growing circle of spam affected users. Uncontrolled and manual filtering, on the other hand, is impossible. Numerous businesses and institutions have made sustained efforts to eradicate this phenomenon; numerous techniques have been proposed, and 3 ITM Web of Conferences 42, 01001 (2022) ICACS21 numerous studies have been conducted on email spam filtering techniques. Additionally, these organizations sought to ascertain the economic impact and disadvantages of the email service on its recipients. These efforts resulted in the discovery of numerous studies and numerous strategies, which have been implemented on a variety of datasets to achieve the desired result.

As a conclusion to this section, prior research and related work demonstrate that: there are numerous email spam filtering techniques, each with its own set of advantages and disadvantages. These techniques vary in their approach to remove or reduce spam. Some are dependent on the user, such as the community blacklists technique while others operate independently of the end user, such as the Bayesian analysis technique. Multiple

studies indicate that using decision tree algorithms to classify email text is quite

effective and efficient at detecting spam. Therefore, we set out to find the best decision tree algorithm for spam email classification.

VII. CONCLUSION

Email has been the most important medium of communication nowadays; through internet connectivity any message can be delivered to all over the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank, related to money or anything that causes destruction to single individual or a corporation or a group of people. Besides advertising, these may contain links to phishing or malware hosting websites set up to steal confidential information. Spam is a serious issue that is not just annoying to the end-users but also financially damaging and a security risk. Hence this system is designed in such a way that it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message which would be of great benefit to individuals as well as to the company .In the future this system can be implemented by using different algorithms and also more features can be added to the existing system.

REFERENCES

[1] Aggarwal, C. C., & Reddy, C. K. (2013). Data clustering: algorithms and applications. Boca Raton.

[2] FL: CRC Press. Arribas-Bel, D., & Hajkowicz, S. A. (2018). Big data for climate change and disaster resilience: challenges and opportunities.

[3] 3) International Journal of Digital Earth, 11(3), 228-245. Castillo, C., & Huerta, R. (2018). Big data in weather forecasting.

[4] In Handbook of Big Data (pp. 697-717). Springer. Chen, M., Mao, S., & Liu, Y. (2014).

[5] Big data: a survey. Mobile Networks and Applications, 19(2), 171-209. Doucet, R., & Gauthier, S. (2018).

[6] Integrating big data and GIS for climate change and renewable energy planning: A systematic review. Renewable and Sustainable Energy Reviews, 82, 460-472.

[7] Hara, T., Nakamura, T., & Ohta, T. (2016). A big data analysis approach to climate change impact assessment.

[8] Environmental Modelling & Software, 78, 128-139. Huang, Y., & Ji, Y. (2017).

[9] A review of data mining and big data analytics for climate and environment. Journal of Meteorological Research, 31(1), 1-19. Iacobacci, M., & Nardi, D. (2019).

[10] Big data and artificial intelligence in meteorology and weather forecasting. ArXiv, abs/1903.08117. Kitchin, R. (2014).