How to Cite:
Soms, N., Hariharan, S., Jeeva, K., & Karthick, C. (2022). Naive bayes machine learning framework for auto detection of spam mails. *International Journal of Health Sciences*, *6*(S3), 7270–7277. https://doi.org/10.53730/ijhs.v6nS3.7742

Naive bayes machine learning framework for auto detection of spam mails

Nisha Soms
Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore – 641010
Email: nishasoms.cse@srit.org

Hariharan S.
Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore – 641010
Email: hariharan.1802030@srit.org

Jeeva K.
Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore – 641010
Email: jeeva.1802037@srit.org

Karthick C.
Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore – 641010
Email: karthick.1802043@srit.org

***Abstract*---**Nowadays, Email spam is a major problem, with the increase usage of internet for product promotions, e-banking, e-loans, health related articles, movie news etc to users. Besides the people are practicing unwanted and unethical conduct as in phishing and mailing fraudulent information. Sending suspicious links through spam emails can harm our system and can also be sought after our system. Creating a fake user identity and email account is such easy for spammers. They pretend like a genuine entity in their spam emails and target those people who are not aware of these frauds. So, it is essential to identify those spam emails which are fraud. In this proposed work, we have applied machine learning algorithm to identify whether the incoming mail is spam or ham. This project will discuss how the Naive Bayes machine learning algorithm is applied on our data sets and how it has been selected for the email spam detection with the best precision and accuracy.

***Keywords*---**spam, ham, email, naïve bayes, machine learning.

Introduction

Nowadays we are living in a digitized world where technology overrules and makes our everyday tasks easier. In other words, we are in a technical world and it has become a fundamental need in our lives. Day by day a remarkable internet revolution had brought up numerous technologies into this globalized world which resulted in countless communication from one end to another end, thereby creating tons of data just by emails. It has become the second nature for more people. While e-mails are necessary for one and all, they also appear with redundant, unavoidable bulk mails, which are also referred to as spam mails. Anyone who access the internet tends to receive the spam on their devices. Most of the spam emails deflect people's concentration away from actual and noteworthy emails and mistakenly lead them towards susceptible situations. Spam emails fills up inboxes or storage capacities, causing to slow down the pace of the internet to a larger level [10].

The spam e-mails comprise the capacity of corrupting one's system by injecting viruses into it, or by sneaking into useful information and stealing from it. These unwanted actions persuade people to be victims of threat to save their own data. The identification of spam emails is a very tedious and boring task and can get frustrating sometimes if done manually. Hence spam detecting software is needed to auto detect spam, which could save some time. To solve this problem, various spam detection techniques are used nowadays. The most widespread technique for spam detection is the utilization of the Naive Bayesian method which feature sets to assess the existence of spam and ham keywords.

Background history

Email spam also referred to as junk emails or spam emails are unremitting messages sent in bulk via email. The process is called spamming. The name comes from a Monty Python sketch in which the name of the canned pork product Spam is omnipresent, unavoidable, and repetitive. Email spam had stable grown since the 90s, and by 2014 it was estimated to account for around 90% of the total email traffic. At the cost of recipient, it is effectively postage due advertising. This makes it an excellent example of a negative externalize [8]. Most email messages are in the form of commercial advertisements. Whether commercial or not, many are not only exasperating as a form of attention theft, but also dangerous because they may contain links that would lead to phishing websites or sites that are hosting malware or might include malware as file attachments. Spammers gather email addresses from customer lists, chat rooms, newsgroups, websites, and viruses thus producing address book. These collected email addresses are used by spammers [9]

Problem statement

The tough competition between filtering method and spammers is going on further, as spammers have begun to use inappropriate methods to overcome the spam filters like using random sender by the addresses or appended random characters at the beginning of mails or end of the emails subject line. Spams are mails which are totally a waste of time to the user. They have to manually sort the

unwanted junk mail which consumes storage space and communication bandwidth. The existing algorithm must be constantly updated with datasets [13] in order to maintain the efficiency of the Model.

Applications
Virus protection

Spam filters can be used to protect the system from exposure to a virus which manipulates in computer upon opening spam emails.

SMS Spam Detectors

The machine learning model built can classify the spam short message services from genuine messages.

Scope

The scope of the proposed work is to construct a machine learning model for classifying the spam emails from the legitimate mails received. The model is built using Multinomial Naïve Bayes and streamlit library in python is used for creating a front end. Usage of Multinomial Naïve Bayes helps us in calculating the probability faster compared to other Naïve Bayes algorithms. The main objectives are to classify spam emails and preventing spam emails being opened. The designed model also has a customization feature where the user himself may provide a keyword for which he is not interested to receive emails from or about it.

Literature survey

Alexy Bhowmick, Shyamanta M. Hazarika et al gifted a wide ranging appraisal about the foremost effectual content-based e-mail spam filtering method. They focused wholly on Machine Learning-based spam filters and its variants. They prepared an extensive analysis ranging from measuring the appropriate ideas, efforts, usefulness, and therefore the current progress. The prerequisite explanation of the setting examines the fundamentals of e-mail spam filtering, the growing nature of spam, how spammers are connecting with the E-mail Service Suppliers (ESPs), and mechanism of how the Machine Learning distinguishes spam. They also covered the effect of Machine Learning-based filters and explored its cases for post developments [4]. S.Ananthi and Dr. S.Sathyabama et al performed a survey on spam filtering techniques, explored the utilization of k-Nearest-Neighbor algorithmic program because it forms the basis for customized spam filters. Many alternative classifiers like Naive theorem classifier, Random Forest Tree, and Heuristic rules square measure combined to construct hybrid spam filtering systems to improve the performance of classification. At constant time, significant experiments on square measure was performed in preprocessing steps to check their impacts on the constant algorithmic program and the results did targeted spam filtering based on k-NN algorithmic program [3].

Vinita Shah of Iran and Patel Bhargesh et al (2018) explored the trend to recommend a novel spam detection procedure, termed as, the text cluster supported vector area model. The authors explains this model as a system which

mechanically constructs the spam detection model by making use of contents of assorted mail forms and determines spam with competence. To build the spam detection model, they apply the cluster algorithm referred to as the spherical suggesting algorithmic program which suits best for every incoming mail. This algorithmic program divides the mails into a predefined range of clusters. For every cluster, a cluster center of mass vectors square measure is found as the Cluster Representative. By getting the clusters, a similarity calculation is carried out between a brand new mail and training email datasets. Hence the text cluster supported vector area model is proved to be successful. However, the contents of assorted forms of mail square measure highlight many term data as the center of mass vectors [5].

K Sai Prasanthi, T Deepika, S Anudeep, M Sai Koushik et al (2019) described the Support Vector Machine (SVM) Learning algorithm. Support Vector Machine is employed for categorization and conjointly works for regression issues whenever the knowledge sets square measure would not train the SVM to classify any new data that it receives. SVM is a supervised machine learning algorithm designed to get a hyper plane thereby classifying the dataset [13] into totally different categories. The SVM maximizes the gap between totally different categories which are attributable to the existence of the many linear hyper planes, termed as margin maximization [6].

Megha et al (2019) in their article experiments and ensures that the Naive Bayes classifier is a straightforward probabilistic classifier with sturdy supposition of independence. In other words, the Naive Bayes classifier assumes the existence or lack of a specific property of a category when it is not associated with the presence or absence of another feature, considering their category variable as a category chance Model while training during a supervised learning setting. A bonus of the naive Bayes classification is that it needs solely low quantity of coaching knowledge to predict the factors needed for its classification. The theorem presumes that the information fits into a specific class. The authors had presented a competence to the then calculated chance turning that the idea is true [1]. MadhuryaT and KarthikV emphasized that E-mail is one amongst the foremost wide used modes of written language over the past 20 years through the net, conjointly the one amongst the best ways of communication that has been accepted for personal message or vocation purpose of communication and its traffic has enhanced sharply with the looks of World Wide internet. Thus these days, everybody have a minimum of one e-mail account. Generally user receives the e-mail consisting of same content repeatedly from multiple users [2].

M. Ramprasad et al, in their paper, presented how the spammers fills our inbox with many extraneous emails causing harmful effects as in stealing our sensitive data like files and contacts from our local device. No matter however we have the newest technology, it is tough to notice spam emails. The agenda of their paper puts forward a Term Frequency Inverse Document Frequency (TFIDF) approach practiced by implementing the Support Vector Machine algorithm. The results square measures were compared in terms of the confusion matrix, accuracy, and exactness. This approach offered accuracy up to 99.9% based on training knowledge and 98.2% based on testing [7].

Existing and proposed system
Existing System

In the existing system, due to rising figures of email users, the growing quantity of the spam emails generated show concerns. It is currently becoming even tougher to handle a variety of e-mails for data processing and machine learning. For that reason, lots of developers have applied comparative studies to various classification algorithms in its execution and their results lead towards categorizing emails effectively by making use of performance metrics. Consequently, it is better to identify an algorithm that produces the effective conclusion for any explicit metric so that the correct categorizing of emails and spam or ham could occur.

Proposed System

In proposed method, we used Multinomial Naïve Bayes algorithm to categorize the spam emails from legitimate ones. The primary step is to select the data set file and apply feature extraction technique for extracting its features, for which we apply the Word count algorithm. After that, preprocessing of the dataset is required. The dataset contains data that are extracted from various sources and forms the main characteristic feature of the extraction to be worked upon. In the formation of data, we have to predict the probability of spam and not spam words in the document. Finally, the next step is to test the data with the help of Naïve Bayesian Classifier which calculates the probability of spam and non-spam mail causing a prediction. If spam words are larger than the words that are not spam in an email, the chosen mail is categorized as spam e-mail. This project needs a coordinated scope of work and they are:

- Modified existing machine learning algorithm.
- Classify the data set after data preprocessing stage followed by data preparation, classification and visualization.
- Calculate the score of data to find out the accuracy of detected spam.

The system overview is presented in Fig.3.1.

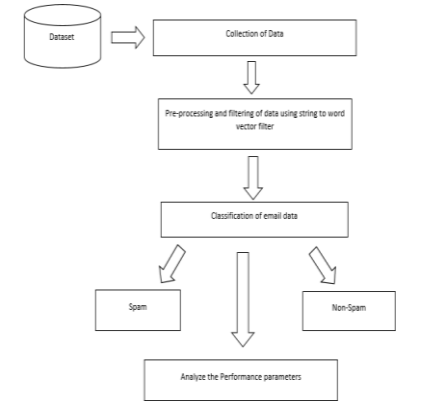

Figure 3.1 System Overview

Methodology

- Dataset processing: The dataset collected is subjected to Data cleaning are followed by exploratory data analysis and text pre-processing.
- Data cleaning: Irrelevant data in the data set and the null values are removed.
- Exploratory Data analysis: the combination of spam and ham messages in the dataset is counted to calculate the prior probability
- Pre-processing: Lower case conversion, Tokenization, Removing special characters, Removing stop words and Punctuation Stemming are performed.
- Selecting algorithm: k-neighbours classifier, Multinomial Naïve Bayes, Decision Tree Classifier, Logistic Regression, Random Forest Classifier are compared by training them. Based on the test data, the precession and accuracy scores obtained from each machine learning algorithms are compared.

Based on results, the Multinomial Naïve Bayes is found best among the algorithms and hence the proposed model is built using Multinomial Naïve Bayes algorithm. The front end is developed in such a way that it accepts both the single and multiple messages. An additional feature called Spam Customization is also developed.

Spam Customization

The Spam Cutomization feature enables the user himself to mention words for which he is not interested to receive emails about or from such sources. For enabling this feature, a dropdown is designed which contains categories and words related to the dataset or email field. These categories and words are stored in a list. The incoming message is tagged as a spam message if it contains any words related to the field selected in the dropdown and an alert 'related to your Interests' with warning of spam message may be given to the user, if he prefers. Some sample outputs of this work are depicted in Fig 4.1, Fig. 4.2 and Fig 4.3 respectively

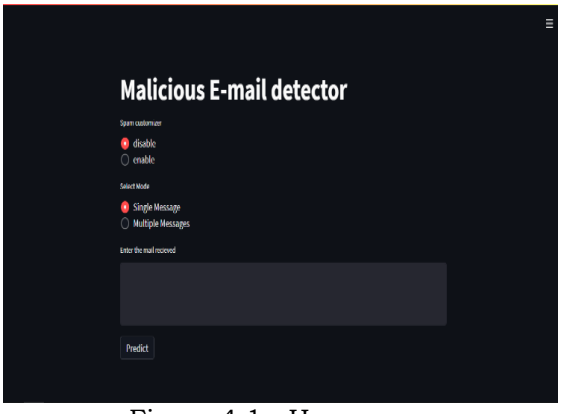

Figure 4.1. Home page

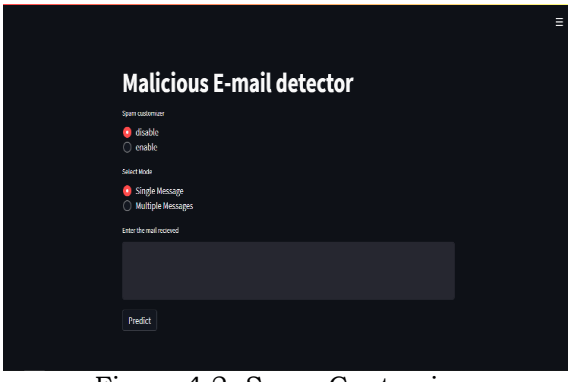

Figure 4.2. Spam Customizer

Figure 4.3. Default Classifier

Conclusion

In this work, the analysis of scientific literature was performed, the characteristics of the malicious user were known. Two information sets were collected for the study, comparable to e-mails and usernames. Next, comparison study among various machine learning classification algorithms was performed to spot malicious users effectively. The Multinomial Naïve Bayes Algorithm is considered to be the best and is applied for classifying ham emails from spam emails. Separate software was built for the same and spam customization feature is introduced. The results disclosed that the classification of emails was effectively done with an accuracy of 99%.

References

1. Megha Bhimashankar Tope, “Email Spam Detection Using Naive Bayes Classifier”, International Journal of Science & Engineering Development Research (IJSDR),Volume 4 ,Issue 6, June 2019
2. Madhurya T1 Karthik V2 ,“Survey on the Content Based Classification of E-mails using Classification Techniques” Vol. 7, Issue 03, IJSRD 2019
3. S.Ananthi Dr. S. Sathyabama ,“spam filtering using k – nn”, Vol-II, No.3, July-Sep 2009

4. Alexy Bhowmick Shyamanta M. Hazarika,“Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends”, June 2016
5. Vinita Shah, Patel Bhargesh “A survey of clustering approaches for spam email detection” , June 2018
6. K sai Prasanthi, T Deepika, S Anudeep, M Sai Koushik “An Efficient Email Spam Detection using Support Vector Machine” Volume-9 Issue-2, December 2019
7. M. Ramprasad1 , N. Harith Chowdary2 , K. Jaswanth Reddy2 and Vishal Gaurav2 “Email spam detection using python & machine learning”2012
8. Rebecca Lieb (July 26, 2002). "Make Spammers Pay Before You Do". The ClickZ Network. Archived from the original on 2007-08-07. Retrieved 2010-09-23.
9. Justin M. Rao & David H. Reiley, 2012. "The Economics of Spam," Journal of Economic Perspectives, American Economic Association, vol. 26(3), pages 87-110, summer.
10. Manoj Sethi1, Sumesha Chandra2, Vinayak Chaudhary3, Yash4, “Email Spam Detection using Machine Learning and Neural Networks” International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 04 | Apr 2021
11. Emmanuel gbenga dada ,joseph stephen bassi,haruna chiroma june-2019-“Machine learning for email spam filtering” Volume 5, Issue 6, June 2019
12. Michael crowford, joseph d.prusa- 5 nov 2015-“Survey of review spam detection using machine learning techniques”, Published: 05 October 2015
13. Dataset source https://www.kaggle.com/venky73/spam-mails-dataset