Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, January-June: 2024

SPAM DETECTION IN EMAILS WITH ADVANCED NATURAL LANGUAGE PROCESSING

Dr. M. Saravanan[1] , T. Saikrishna[2]

[1] Professor, Department of Computer Applications, Aurora's PG College (MBA), Uppal, Hyderabad
Email**:** m.saravananme@gmail.com

[2]Assistant Professor, Department of Computer Applications, Aurora's PG College (MBA), Uppal, Hyderabad
Email**:** thotasai1216@gmail.com

ABSTRACT

The problem of spam emails has long plagued computer security. They are exceedingly risky for computers and networks and very expensive economically. Despite the rise of social networks and other online platforms for knowledge sharing. The demand for better spam filters has become critical as a result of the growing reliance on email communication over time. There is a dearth of study on text modifications, despite the fact that numerous spam filters have been developed to help stop these spam emails from getting into a user's mailbox. Naive Bayes is currently one of the most often used spam classification techniques due to its effectiveness and simplicity. Additionally highly accurate is Naive Bayes. For this reason, we used the Naive Bayes Spam Filter in this project so that it could identify textual changes and appropriately categorize emails as spam or ham. Our technique improves the accuracy of Naive Bayes by more than 200 percent when compared to Spam assassin by combining semantic-based, keyword-based, and machine learning algorithms.

Keywords: Spam Detection, Emails, Natural Language Processing

1 INTRODUCTION

Spam is a term used to describe unsolicited emails that seek to manipulate the recipient or that merely indiscriminately clutter the mailbox. It clogs Internet users' inboxes and is also known as junk mail. Spam emails nowadays come in many different forms, from advertisements to business promotions to dubious product listings to services that seem unpleasant. As a result, it is challenging to distinguish between emails that are spam and those that are not. Usenet, also known as User Network, is an email service that disseminates group discussions or emails targeted at a specific group of individuals connected to a certain service or product. These emails are primarily educational but can clog up a user's mailbox. Netnews is the term for the information that is distributed over the Internet; a collection of these resources intended to convey information on a certain subject is referred to as a "newsgroup." Spammers' main goal is everyone who reads such news from these newsgroups. These news groups are used by spammers to advertise specific irrelevant posts or advertisements. Usenet spam promotes irrelevant posts, depriving users of the newsgroups' usefulness.

Project's Relevance

Email usage has increased as communication becomes more digitalized; in 2016, there were an estimated 2.3 million users of the service. There were 205 billion emails sent and received every day in 2015. By 2019, that number is predicted to have increased to over 246 billion, growing at a rate of 3% annually. Due to the lack of an accurate spam classifier in existing spam detection technologies, the boom in emails has also resulted in an unprecedented rise in spam, accounting for 49.7% of all sent emails. Spam is a concern because spam emails hog processing power, storage space, and network traffic in addition to frequently being the transmitter of viruses. Additionally, because spam reduces productivity at work and costs money, the commercial sector has a strong interest in spam identification. According to estimates, American businesses and consumers lose $20 billion a year, despite the fact that private companies continue to invest in anti-spam software. However, spam advertising brings around $200 million annually. The dynamic nature of spam means that many of the filters in use today, despite years of intensive work toward their enhancement, have limited effectiveness.

2.LITERATURE SURVEY AND RELATED WORK

Swarm Optimization

- Naïve Bayes algorithm is a Bayes theorem based statistical machine learning based approach having properties of strong independence, probability distribution and ability to handle large datasets.
- In NB, probability distribution is evaluated from the frequency distribution of dataset.
- Particle Swarm Optimization (PSO) is swarm intelligence based concept derived in 1995 by Eberhart and Kennedy
- PSO work on the property of stochastic distribution and initially find the local search solution, then individual particle share their solution and global solution is obtained.
- NB having probability distribution property determines the possible class for the email content from the spam class or non-spam class on the basis of keywords present in the email textual data.
- PSO is used to further optimize the parameters of NB approach to improve the accuracy, search space and classification process.

Support Vector Machine

- This paper uses Support Vector Mechanism algorithm to identify spam emails.
- Descriptions as provided on Spam Assassin website for the dataset used in this paper.
- SVM is also considered as an important kernel method, which is one of the most important areas in machine learning concepts.
- Smart Traffic Control System with Application of Image Processing Techniques
- In this work they have also compared Linear and Gaussian as two of the very popular kernel and employed them for the problem of email spam detection
- The two models have been proposed, trained and tested using popular and often used standard database

3 EXISTING SYSTEM

Due to the increase in the number of email users, and at the same time increasing the spam emails in inbox although increases the data storage. So sometimes systems or mobiles are working to slow. This is problems are rectifies some of algorithms.

DISADAVANTAGES
1. Unwanted e-mails irritating internet connection
2. Critical e-mail message are missed and / or delayed.
3. Spam can crash mail servers and fill up hardware

4 PROPOSED WORK AND ALGORITHM

The proposed system is classified to the spam or ham messages in inbox automatically. Once filter the spam mails automatically move the spam folder.

5 METHODOLOGIES

MODULES

1. Load Dataset:
Load data set using pandas read_csv () method. Here we will read the excel sheet data and store into a variable.
2. Split Data Set:
Split the data set to two types. One is train data test and another one is test data set. here we will remove missing values from the dataset.
3. Train data set:
Train data set will train our data set using fit method. 80% of data from dataset we use for training the algorithm.
4. Test data set:
Test data set will test the data set using algorithm. 20% of data from dataset we use for testing the algorithm.
5. Predict data set:
Predict () method will predict the results. In this step we will predict the ranking of the google play store app.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, January-June: 2024

6 RESULTS AND DISCUSSION

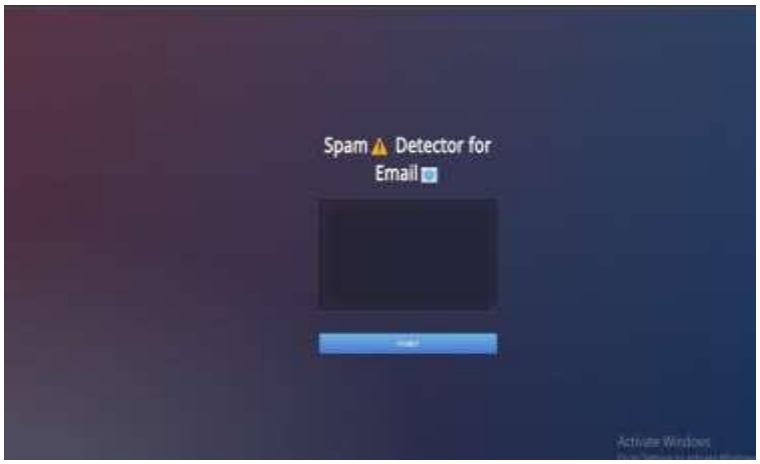

Fig 1: HOME SCREEN DETECTOR OF EMAIL

Fig 2: DETECTOR FOR OTHER EMAILS

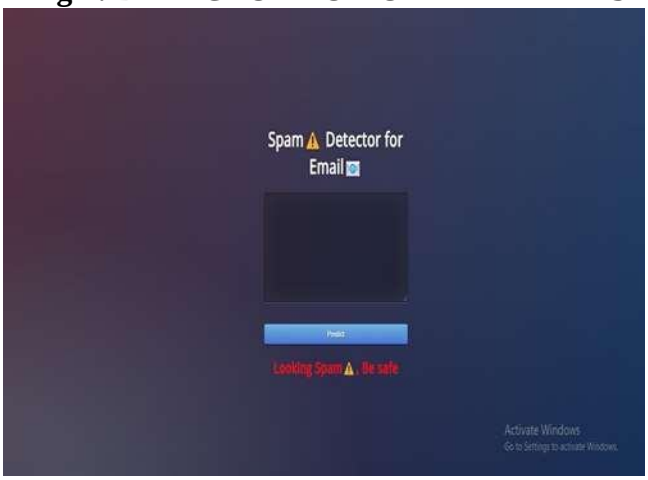

Fig 3: PREDICTED EMAILSPAM

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, January-June: 2024

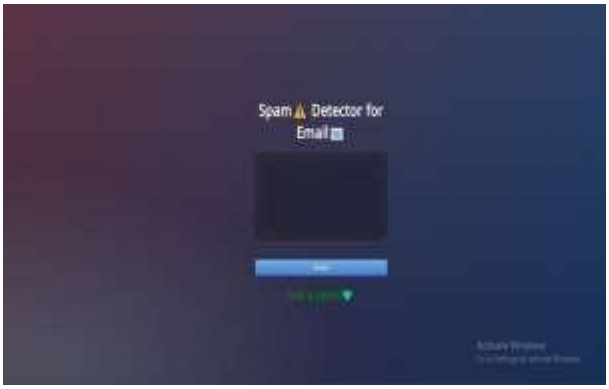

Fig 4: PREDICT HAM MAIL

6. CONCLUSIONAND FUTURE SCOPE

The Naive Bayes Spam Filter is what we suggested. The Naive Bayes Classifier was enhanced by the implementation of the method. Naive Bayes is a good choice for real-time spam filtering because of its lightning-fast processing speed and tiny training set capacity. Additionally, we use the Intelligent Text Modification approach to recognize texts that contain diacritical marks and leetspeak. Emails can be categorized as either spam or ham.through the development of a Naive Bayes Classifier augmentation. Because of the new addition's strong recall and precision rates, we also discovered that ham classification was improved. We showed that the number of spam emails that were mistakenly identified as ham emails was continuously decreased by our algorithm.

Future Scope of the Project:

Future Scope of the Project: We want to develop an API for the same and test it in an actual setting in the future. Our goal is to optimize this project for a significantly larger dataset. We plan to apply our modification to additional machine learning spam filters, including vector space models, clustering, and artificial neural networks, as it successfully improves the Naive Bayes spam filter. By combining these additional techniques, spam detection will be improved across a wide range of systems, leading to the eventual creation of a sophisticated spam detector for text alterations. We want to develop an API for the same in the future and test it in an actual setting. Our goal is to optimize this project for a significantly larger dataset. We plan to apply our modification to additional machine learning spam filters, including vector space models, clustering, and artificial neural networks, as it successfully improves the Naive Bayes spam filter. By combining these additional techniques, spam detection will be improved across a wide range of systems, leading to the eventual creation of a sophisticated spam detector for text alterations.

8 REFERENCES

1. Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization
2. Email Spam Classification by Support Vector Machine
3. Intelligent Model for Classification of SPAM and HAM
4. Team, Radicati. "Email Statistics Report, 2015-2019. The Radicati Group." (2015).
5. Androutsopoulos I., J. Koutsias, K. V. Chandrinos, G. Paliouras,and C. D. Spyropoulos, "An evaluation of naive bayesian antispam filtering", In: 11th European Conference on Machine Learning, pp.9-17, Barcelona, Spain, 2000.
6. GitHub, Inc, "Spam Assassin," 21 April 2016. [Online]. Available: https://github.com/dmitrynogin/SpamAssassin.git. [Accessed 20 August 2017].