

EMAIL SPAM FILTER USING NATURAL LANGUAGE PROCESSPADALI VIMALA¹, B.S.MURTHY²**1. PG student, D.N.R. COLLEGE, P.G. COURSES (AUTONOMOUS), BHIMAVARAM-534202.****Email id:-vimala84861@gmail.com****2. Assistant Professor in DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS, BHIMAVARAM-534202.****Email id: - suryanarayanamurthy.b@gmail.com****ABSTRACT**

Spam emails have been a chronic issue in computer security. They are very costly economically and extremely dangerous for computers and networks. Despite of the emergence of social networks and other Internet based information exchange venues. Dependence on email communication has increased over the years and this dependence has resulted in an urgent need to improve spam filters. Although many spam filters have been created to help prevent these spam emails from entering a user's inbox, there is a lack of research focusing on text modifications. Currently, Naive Bayes is one of the most popular methods of spam classification because of its simplicity and efficiency. Naive Bayes is also very accurate. Thus, in this project we implemented the Naive Bayes Spam Filter so that it can detect text modifications and correctly classify the email as spam or ham. Our algorithm combines semantic based, keyword based, and machine learning algorithms to increase the accuracy of Naive Bayes compared to Spam assassin by over two hundred percent

1 INTRODUCTION

Spam refers to an email aimed manipulating an individual to whom it is aimed at or just randomly flooding the inbox. It is also called as junk mail and it floods Internet clients Inboxes. Today spam emails are of a variety of types ranging from ads to business promoting to doubtful products to some objectionable services.

Therefore it is difficult to identify and classify an email as spam or non-spam.

Usenet also called as User Network is an email service that distributes group talks or emails aimed at a particular group of people associated with a certain service or product and are mostly informative but do crowd up the inbox of the user. The data that goes over the Internet is called Netnews" an accumulation of these data that is aimed at providing message about a specific topic is called a "newsgroup". People that read such news from these newsgroups are the prime target of Spammers. Spammers use these news groups for the promotion of certain unrelated ads or unrelated posts. Usenet spam robs clients of the utility of the newsgroups by promoting other unrelated posts.

1.1 Relevance of the Project

As the digitization of communication grows, electronic mail, or emails, has become increasingly popular; in 2016, an estimated 2.3 million people used email. In 2015, 205 billion emails were sent and received daily, which is expected to grow at an annual rate of 3% and reach over 246 billion by 2019. However, the growth in emails has also led to an unprecedented increase in the number of illegitimate mail, or spam - 49.7% of emails sent is spam - because current spam detection methods lack an accurate spam classifier. Spam is problematic not only because it often is the carrier of malware, but also because spam emails hoard network bandwidth, storage space, and computational power. Additionally, the commercial world has significant interests in spam detection because spam causes loss of work productivity and financial loss. It is estimated that American firms and consumers lose 20 billion annually, even while sustained by the private firms' investment in anti-spam software. On the other hand, spam advertising earns 200 million per year. Although extensive work has been done on spam filter improvement over the years, many of the spam filters today have limited success because of the dynamic nature of spam.

2. LITERATURE SURVEY AND RELATED WORK**2.1 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.

2.2 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.

DISADVANTAGES

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.

6 RESULTS AND DISCUSSION

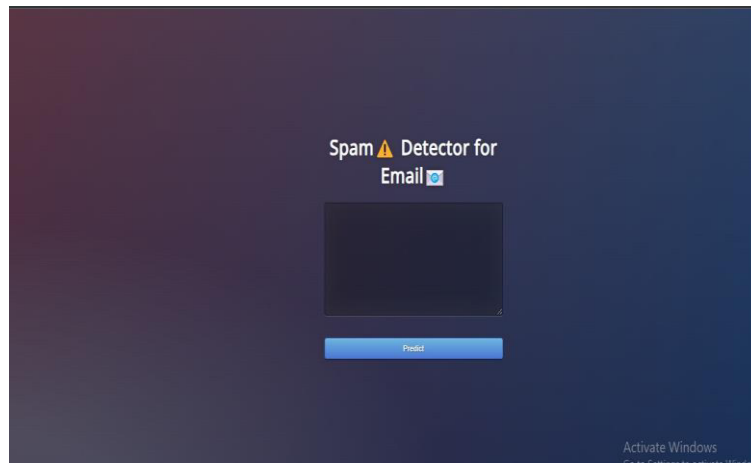


Fig 1: HOME SCREEN DETECTOR OF EMAIL

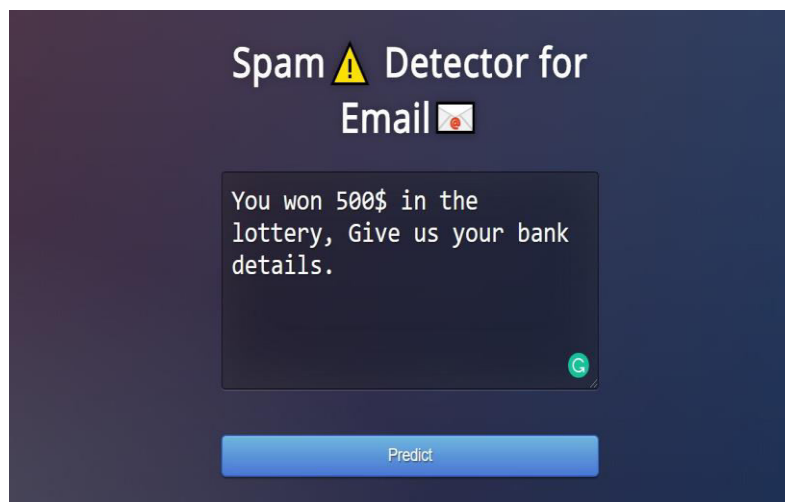


Fig 2: DETECTOR FOR OTHER EMAILS

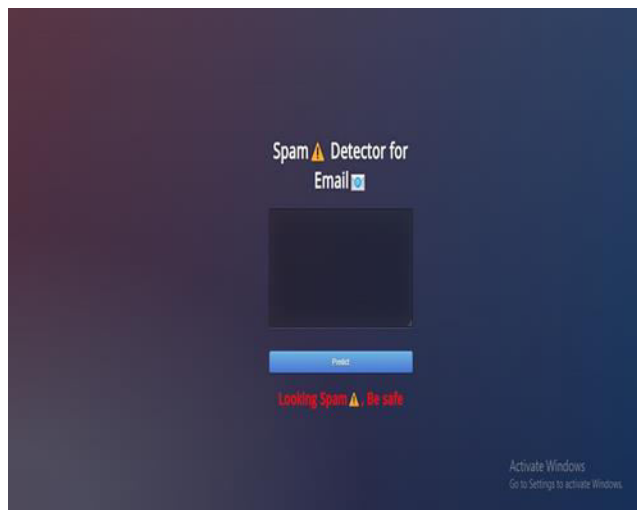


Fig 3: PREDICTED EMAILSPAM

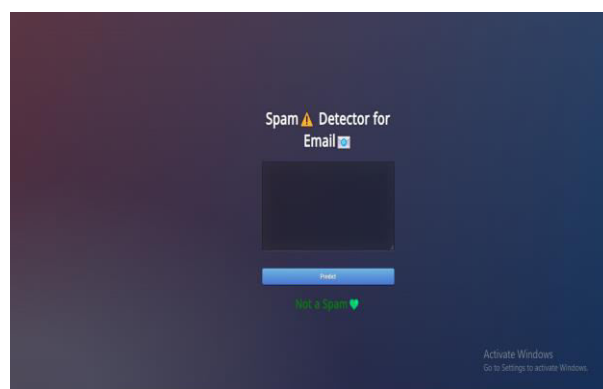


Fig 4: PREDICT HAM MAIL

7. CONCLUSION AND FUTURE SCOPE

We proposed the Naive Bayes Spam Filter. The algorithm was implemented as an enhancement for Naive Bayes Classifier. Naive Bayes has a very fast processing speed and allows for a small training set, hence is suitable for real-time spam filtering. We are also using Intelligent Text Modification method to identify messages containing leetspeak and diacritic. We are able to classify email as spam or ham .By creating an addition to Naive Bayes Classifier. We also found that our new addition helped improve ham classification due to the high recall and precision rates. We demonstrated that our algorithm consistently reduced the amount of spam emails misclassified as ham email.

Future Scope of the Project:

In the future we would like to create an API for the same and test that in real world environment. We will try to optimize this project for much larger amount of dataset. Since our addition successfully enhances the Naive Bayes spam filter, we will try to implement the addition onto other machine learning spam filters such as Vector Space Models, clustering, and artificial neural networks. Combining these other methods will allow the improvement of spam detection across many different systems to ultimately create a well-developed spam detector for text modifications. In the future we would like to create an API for the same and test that in real world environment. We will try to optimize this project for much larger amount of dataset. Since our addition successfully enhances the Naive Bayes spam filter, we will try to implement the addition onto other machine learning spam filters such as Vector Space Models, clustering, and artificial neural networks. Combining these other methods will allow the improvement of spam detection across many different systems to ultimately create a well-developed spam detector for text modifications.

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].