IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EMAIL SPAM DETECTION USING KNN ALGORITHM

[1]Dr. M. Senthil, [2]D. Sruthi, [3]G. Pravallika, [4]Ch. Ajay, [5]A. Kumar

[1]Professor, Department of Computer Science and Engineering, QIS college of Engineering and Technology,Ongole, Andhra Pradesh, India
[2,3,4,5]Student, Department of Computer Science and Engineering, QIS college of Engineering and Technology ,Ongole, Andhra Pradesh, India

Abstract: Spam filtering technology must be developed quickly due to the increase of unsolicited emails, or spam mails. Computer security has struggled with spam emails on a consistent basis. They are incredibly expensive economically and exceedingly risky for networks and computers. Spam emails are found and filtered using machine learning techniques.
This project mainly focus on machine learning used to find and remove spam emails. Using the K-nearest neighbor algorithm for email spam detection is one of the simple supervised learning techniques.
Applications are utilizing machine learning techniques in spam filters for email in Gmail and Outlook, two of the top internet service providers. Regression method used to detect and filters spam mails.

Index Terms - Super vised learning, spam or ham, tokenization, lemmatization, stemming, K-nearest neighbor.

I. INTRODUCTION

Spam is information that is intended to be distributed to many people. The automated detection of spam is done by a spam filter. Spam inhibits the user from efficiently using their time and storage. Users who receive span mails find it very irritating. Scams and other fraudulent practices of spammers with the effect of sensitive personal information like passwords, credit card numbers. The spammer is the individual who sends the unsolicited mails. They gather malware and email addresses from many websites. Spam inhibits the user from efficiently using their time and storage. The massive amount of spam emails travelling across computer networks has a negative impact on email servers' memory, CPU, and user time. Google machine learning model has now developed to the point where it can reliably identify and filter out spam and phishing emails with a 99.9% accuracy rate. This suggests that one thousand emails every day manage to slip past their spam filter. Between 50 and 70 percent of the emails that Gmail receives are unsolicited, according to Google's statistics. Spam email volume decreased to 49.7%, and by July 2015, it had further decreased to 46.4%, according to Symantec, a maker of antivirus software. For the first time since 2003, the spam email percentage dipped in below 50%. Kaspersky Lab found between 3 million and 6 million in 2015 as of June. 22,890,956 spam emails were uncovered by Kaspersky Lab March 2016. According to statistics, spam accounts for 56.87% of all global email traffic. China Net, Amazon, and Airtel India are the three networks housing the most spammers as of December 13, 2021.

Machine learning algorithm include K-nearest neighbor algorithm used to detect which mail is spam and which mail is ham.

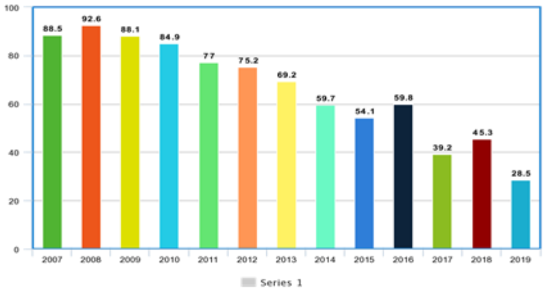

FIGURE 1: Spam Infographics

II. EXISTING SYSTEM

The initial transformation begins with pre-processing activities such as data extraction, classification of email content, and process analysis. The data is split into two sets using a vector expression.
To detect whether an email is spam or not, on training and test data sets, machine learning is employed.
20% of the original dataset is utilized to test the model, while the remaining 80% is used for training.

III. PROPOSED SYSTEM

We proposed a system that can be useful not only for our ug project but in real time scenario.
The aim of the project is to make spam mails using the features present in the data set. The dataset is extracted features using the K-Nearest Neighbor technique to create features using python as we can detect and filter spam mails. This extracted data will be using in Regression model. The model formed in scores. The scores classified be spam or not spam. It can predict the output accurately. Its works efficiently on content based emails. It is simple algorithm. This requires high accuracy. Its efficient method for small datasets

IV. IMPLEMENTATION

Our system can be implemented as the following diagram by
1) Data Collection
2) Pre-processing the Data
3) Feature Extraction
4) KNN Implementation
5) Performance Analysis

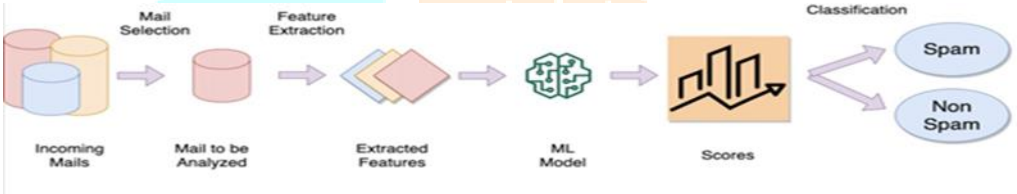

FIGURE 2: System Implementation

V. MODELLING

The below diagram shows that our system It receives messages as input, filters them, and organizes them into inboxes and spam folders which emails are spam. It enters the spam folder. The inbox folder receives the ham mails and stores it there.

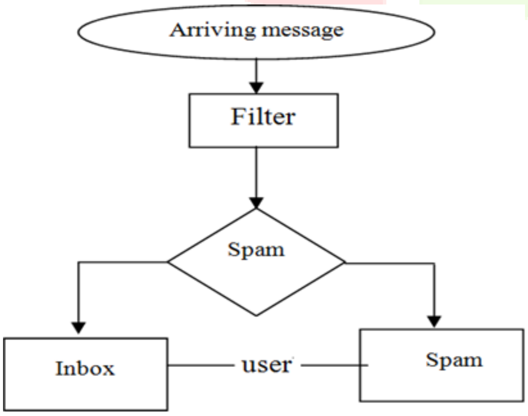

Figure 3: Activity diagram represents of email spam detection

VI. RESULTS

After making our model, we check and validate multiple times and it displaying spam mails for particular webpage, ham mails are displaying particular webpage. Our code to build with python based on the data sets. Finally, it displays using KNN algorithm to classify email into spam and ham like it takes nearest neighbors to detect the mails there.

VII. CONCLUSION

K-nearest neighbor identifies the nearest neighbors like detect the spam mails. It is better performance than other algorithms like super vector machine and naïve bayes algorithms. Spammers are now evolving and sending emails containing pictures and pdf to pass the filter. KNN algorithm is used to determine which emails are spam and which emails are ham messages. The huge body can utilise the recursion method to tell which emails are legitimate and which ones are spam.

VIII. REFERENCES

[1] Detection of Email Spam: A Comparative Analysis of Classification Algorithms, 2018. Osho, O., Ismaila, and Alhassan, along with Shafi'i Muhammad Abdul Hamid, M. S.

[2] Singh, V. K., and S. Bhardwaj's paper, "Spam Mail Detection Using Classification Techniques and the Global Training Set," was published in 2018.

[3] Using evolutionary computation for finding spam patterns from e-mail samples, D. Ruano-Ordas, F. Fdez-Riverola, and J.R. Mendez, 2018.

[4] Shubhangi Suryawanshi, Anurag Goswami, and Pramod Patil are the authors(2019). A Comparative Empirical Analysis of Various ML and Ensemble Classifiers for Email Spam Detection. 69-74. 10.1109/IACC48062.2019.8971582.

[5] Pages 685–690 of The Second International Conference on Intelligent Computing and Control Systems (ICICCS), held in Madurai, India, in 2018. "Email Spam Detection Using Integrated Approach of Nave Bayes and Particle Swarm Optimization," K. Agarwal and T. Kumar.

[6] Award, W.A. and S.M. Elseuofi (2011). Machine Learning Techniques for Email Spam Classification. International Journal of Computer Science & Information Technology, DOI: 3.10.5121/ijcsit.2011.3112.