A CASE STUDY ONE-MALE SPAM DETECTION USING MACHINE LEARNING

Dr.D.Rajeshwari[1], P.Sowjanya[2], Gattu Pavan Kumar[3], Gottemukkula Divya Reddy[4], Balagouni Naveen[5], Dongala Vineeth Reddy[6]

[1]Assistant Professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[2]Assistant professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[3,4,5,6] IV[th] Btech Student, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

ABSTRACT

Email spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. S ending malicious link through spam emails which can harm our system and can also seek in into your system. Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. S o, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning, this paper will discuss the machine learning algorithms and apply all these algorithm on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy.

Key words : Computer Privacy , Computer Security , Deep Learning, Neural Networks, Spam Filtering.

1. INTRODUCTION

Email or electronic mail spam refers to the "using of email to send unsolicited emails or advertising emails to a group of recipients. Unsolicited emails mean the recipient has not granted permission for receiving those emails. "The popularity of using spam emails is increasing since last decade. Spam has become a big misfortune on the internet. Spam is a waste of storage, time and message speed. Automatic email filtering may be the most effective method of detecting spam but nowadays spammers can easily bypass all these spam filtering applications easily. Several years ago, mos t of the spam can be blocked manually coming from certain email addresses. Machine learning approach will be used for spam detection. Major approaches adopted closer to junk mail filtering encompass "text analysis, white and blacklists of domain names, and community-primarily based techniques". Text assessment of contents of mails is an extensively used method to the spams. Many answers deployable on server and purchaser aspects are available. Naive Bayes is one of the utmost well-known algorithms applied in these procedures. However, rejecting sends essentially dependent on content examination can be a difficult issue in the event of bogus positives. Regularly clients and organizations would not need any legitimate messages to be lost. The boycott approach has been probably the soonest technique pursued for the separating of spams. The technique is to acknowledge all the sends other than those from the area/electronic mail ids. Expressly boycotted. With more up to date areas coming into the classification of spamming space names this technique keeps an eye on no longer work so well. The white list approach is the approach of accepting the mails from the domain names/addresses openly whitelisted and place others in a much less importance queue, that is delivered most effectively after the sender responds to an affirmation request sent through the "junk mail filtering system" .

2. LITERATURE SURVEY

Authors have highlighted several features contained in the email header which will be used to identify and classify spam messages efficiently .Those features are selected based on their performance in detecting spam messages. This paper also communalize each features contains in Yahoo mail,Gmail and Hotmail so a generic spam messages detection mechanism could be proposed for all major email providers. In the paper[2], a new approach based on the strategy that how frequently words are repeated was used. The key sentences, those

with the keywords, of the incoming emails have to be tagged and thereafter the grammatical roles of the entire words in the sentence need to be determined, finally they will be put together in a vector in order to take the similarity between received emails. K-Mean algorithm is used to classify the received e-mail. Vector determination is the method used to determine to which category the e-mail belongs to. In the paper[3],authors described about cyber attacks .Phishers and malicious attackers are frequently using email services to send false kinds of messages by which target user can lose their money and social reputations. These results into gaining personal credentials such as credit card number, passwords and some confidential data .In This paper ,authors have used Bayesian Classifiers .Consider every single word in the mail. Constantly adapts to new forms of spam. In the paper[4],proposed system attempts to use machine learning techniques to detect a pattern of repetitive keywords which are classified as spam. The system also proposes the classification of emails based on other various parameters contained in their structure such as Cc/Bcc, domain and header. Each parameter would be considered as a feature when applying it to the machine learning algorithm. The machine learning model will be a pre-trained model with a feedback mechanism to distinguish between a proper output and an ambiguous output. This method provides an alternative architecture by which a spam filter can be implemented. This paper also takes into consideration the email body with commonly used keywords and punctuations. 3 In the paper[5],authors investigated the use of string matching algorithms for spam email detection. Particularly this work examines and compares the efficiency of six well known string matching algorithms, namely Longest Common.

3. SYSTEM IMPLEMENTATION

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system.

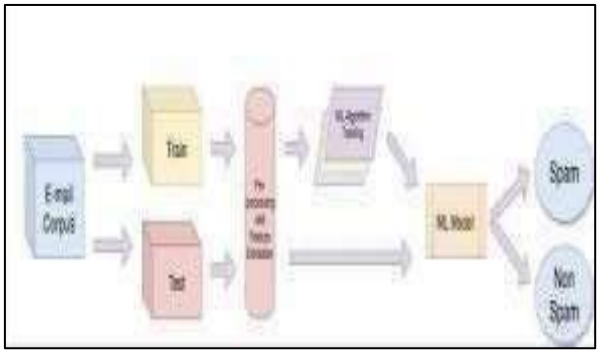

Fig 1 : System Architecture

3.1 SOFTWARE ENVIRONMENT

a. Python

Python is currently the most widely used multi-purpose, is the high-level programming language. Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and the good indentation of the requirement of that the language, makes them readable all the time. Python language is being used by almost all tech-giant companies like Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

b. Machine Learning

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data. 30 Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain.Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

4. RESULTS

With this result, it can be concluded that the Multinomial Naïve Bayes gives the best outcome but has limitation due to classconditional independence which makes the machine to misclassify some tuples. Ensemble methods on the other hand

proven to be useful as they using multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus. Our project, thus spam detection is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria.

5. CONCLUSION

It can be concluded that the Multinomial Naïve Bayes gives the best outcome but has limitation due to class-conditional independence which makes the machine to misclassify some tuples. Ensemble methods on the other hand proven to be useful as they using multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus. Our project, thus spam detection is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria. Therefore, at this it is an only limited body of the email. There is a wide possibility of improvement in our project. The subsequent improvements can be done: "Filtering of spams can be done on the basis of the trusted and verified domain names." "The spam email classification is very significant in categorizing e-mails and to distinct e-mails that are spam or non-spam." "This method can be used by the big body to differentiate decent mails that are only the emails they wish to obtain."

6. FUTURE SCOPE

we reviewed machine learning approaches and their application to the field of spam filtering. A review of the state of the art algorithms been applied for classification of messages as either spam or ham is provided. The attempts made by different researchers to solving the problem of spam through the use of machine learning classifiers was discussed. The evolution of spam messages over the years to evade filters was examined. The basic architecture of email spam filter and the processes involved in filtering spam emails were looked into. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter. The challenges of the machine learning algorithms in efficiently handling the menace of spam was pointed out and comparative studies of the machine learning technics available in literature was done. We also revealed some open research problems associated with spam filters. In general, the figure and volume of literature we reviewed

shows that significant progress have been made and will still be made in this field.

7. REFERENCES

[1] Suryawanshi, Shubhangi & Goswami, Anurag & Patil, Pramod. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74. 10.1109/IACC48062.2019.8971582.

[2] Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. (2019). A Comprehensive Survey for of Intelligent Spam Email Detection.the IEEE Access, 7, 168261-168295.

[3] K. Agarwal and T . Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second of Intelligent Computing and Control

[4] Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "T ext and imagebased spam email classifing using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information T echnology (ICROIT), 2014 International Conference on, pp.153 -155. IEEE, 2014

[5] Mohamad, Masurah, and Ali Selamat. "An evaluation on the efficiency of feature selection in spam email classification." In Computer, the science Communications, and Control T echnology (I4CT), 2015 International Conference on, pp. 227 -231. IEEE, 2015

[6] Shradhanjali, Prof. T oran Verma "EMail Spam Detection and Classification Using SVM and Feature Extraction"in International Jouranl Of Advance Reasearch, Ideas and Innovation In T echnology,2017 ISSN: 2454-132X Impact factor: 4.295

[7] W.A, Awad & S.M, ELseuofi. (2011). Machine Learning Methods for Spam Mail Classification. International.

[8] A. K. Ameen and B. Kaya, "Spam detection in online social networks Mahdian, B.; Saic, S. A bibliography blind methods for identifying image Signal Process. Image Commun. 2010, 389–399.