IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006209|

Email Spam Detection and Filtering Using Machine Learning

Tausif Mansuri, Neeraj Oswal, Prof. M. K. Nivangune, Yogesh Sharma, Taher Patanwala

Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India

Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India

Professor, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India

Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India

Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India

ABSTRACT: Many techniques have been proposed to combat the upsurge in spam. All the proposed techniques have the same target, trying to avoid the spam entering our inboxes. Spammers avoid the filter by different tricks and each of them needs to be analyzed to determine what facility the filters need to have for overcoming the tricks and not allowing spammers to full our inbox. Different tricks give rise to different techniques. This work surveys spam phenomena from all sides, containing definitions, spam tricks, anti spam techniques, data set, etc. Finally, we discus the data sets which researchers use in experimental evaluation of their articles to show the accuracy of their ideas. Machine learning methods of recent are being used to successfully detect and filter spam emails. We present a systematic review of some of the popular machine learning based email spam filtering approaches. Our review covers survey of the important concepts, attempts, efficiency, and the research trend in spam filtering. The preliminary discussion in the study background examines the applications of machine learning techniques to the email spam filtering process of the leading internet service providers (ISPs) like Gmail, Yahoo and Outlook emails spam filters.

KEYWORDS: Computer science, Analysis of algorithms ,Machine learning, Spam filtering, Deep learning ,Neural networks, Support vector machines, Naïve Bayes.

I. INTRODUCTION

For sharing of important and official information, email is used as a default medium of communication. It is used widely as it also provides the confidentiality of the data shared. But with the pros also comes the cons, as many people misuse this reliable and easy way of communication by sending unwanted and useless bulk messages for their own personal benefits. These unwanted emails affect the normal user to face the problems like flooding of the mail box with unwanted emails making it harder to look for the useful once, even sometimes one may skip through important and useful emails because of all these unwanted emails. So, this gives rise to a need of a strong email spam detector which can filter maximum amount of spam emails with a greater accuracy so that a genuine email does not get filtered as spam .In recent times, spam has become a huge problem on the internet. The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time .It is responsible for over 77% of th
whole global email traffic .

A. Motivation

With the internet becoming an integral part of our lives, leading to substantial use of the internet, number of email users is also increasing day by day. With so much increase in number of users. There's an excessive increase in the use of email that has caused problems by unnecessary bulk email messages known as "Spam". Spam emails are formulaic and their tricks are so obvious. It's easy to look at spam and laugh, but the unfortunate reality is that people are still falling for spam. If they weren't, we wouldn't see so much spam. Maybe they fall for the old "Nigerian prince" email and lose money, order some cheap pharmaceuticals of questionable purity, or click a link and download malware. So

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006209|

how would we even begin solving the spam problem? Well, we could pass laws making spam illegal, have legitimate services shut down spammers who use their services, and develop good spam filters to prevent as many spam messages as possible from reaching people's inboxes. These messages feature content that is usually of little interest to majority of the recipients and sometimes they can also contain viruses and so we are here to solve the problem using some machine learning techniques and filters . Being final year engineering students, and the email being an integral part of our lives, motivated us to take this project.

B. Problem statement

1.Problem statement

➢ Email Spam detection and filtering with Naive Bayes Classifier Algorithm .

2.What is Naive Bayes algorithm in short and how will it help?

➢ Spam emails are classified,segregated and sorted using Naive Bayes Classifier. It basically relies on Bayes Rule. This algorithm basically classifies each object according to their features. Bayes Rule below shows us how to calculate the posterior probability for just one feature. The posterior probability of the object is calculated for each feature and then these probabilities are multiplied together to get a final probability. This probability is calculated for the other class as well. Which ever has the greater probability that ultimately determines what class the object is in.

II. BACKGROUND

Recommender systems are designed to recommend things to the user based on an analysis of their data. It has the capability of producing outputs or recommendations by predicting the set of items based on the user's history. These systems generally work by dealing with huge amounts of information based on the data provided by the user along with their preferences. The design of such systems depends on the domain and the particular characteristics of the data available. Recommender systems differ in the way they analyze the data sources to develop affinity between users and items, which can be used to identify well-matched pairs. It is necessary to identify the properties which will help in creating a successful recommender system in the context of a specific application.

There are some popular algorithms used in developing these systems. Collaborative filtering is used to recommend items to the users that other users with similar tastes have liked in the past. Content-based filtering is based on the preference or ratings a user would give an item. It compares the user's preferences to the features of the items. The one having the most overlapping features is recommended to the user. Hybrid filtering is the combination of both Content-based and Collaborative filtering. It makes the predictions of both the algorithms separately and merges them to give more accurate recommendations.

III. RELATED WORK

Spammers are now able to launch large scale spam campaigns, malware and botnets helped spammers to spread spam widely. Upon receiving and opening a spam email, Internet users is exposed to security issues as spams are normally broadcast for bad intention. One of the common email spam example received by users are an email requesting for IDs and passwords(Refer to Figure 1).


From: avoth@cogeco.ca [mailto:avoth@cogeco.ca] On Behalf Of Webmail.Uchicago.edu@cogeco.ca
Sent: Thursday, July 24, 2008 6:46 PM
Subject: Quoting Uchicago.edu, Member.Services@ Uchicago.edu

Dear Uchicago.edu, email account user,
We are currently verifying our subscribers email accounts in other to increase the efficiency of our webmail futures. During this course you are required to provide the verification desk with the following details so that your account could be verified;
CNetID:.....
Password:.....
Territory:.....

Kindly send these details so as to avoid the cancelation of your email account.

Thanks, Uchicago.edu, Team

Figure 1. Sample of spam data requesting for ID and password

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006209|

Figure 1. Sample of spam data requesting for ID and password Several machine learning algorithms have been employed in anti-spam e-mail spam filtering, including algorithms that are considered top-performers in Text Classification [3], like Boosting algorithm, Support Vector Machines (SVM) algorithm [5] and Na¨ıve Bayes algorithm [7]. Konstantin Tretyakov et al., [6] have evaluated several most popular machine learning methods i.e., Bayesian classification, k-NN, ANNs, SVMs and of their applicability to the problem of spam-filtering. In this work, the author proposed most trivial sample implementation of the named techniques and the comparison of their performance on the PU1 spam corpus dataset is presented. The author used extracting feature to convert all messages to vectors of numbers (feature vectors) and then classify these vectors. This is because most of the machine learning algorithms can only classify numerical objects like vector.

IV. SYSTEM IMPLEMENTATION

In this section, we present a brief explanation of the implementation of our proposed system.
We have used the Na¨ıve Bayes algorithm The working of the Na¨ıve Bayes algorithm is explained with the help of the following steps.
Step 1Consider a random email from the link spam dataset for experimentation
Step 2The considered email is in raw form. To perform the feature extraction/selection and classification procedure, initially email is needed to pre-process. Pre-processing involves thesteps of tokenization, stemming and stop word removal.
2.1. Initially, tokenize the email into individual keywords. Tokenization split each individualword into different tokens.
2.2. Remove the stop words from the obtained tokens.
2.3. Perform stemming on the tokens obtained from the previous step. Stemming process reducesthe size of the word to its root word. For stemming, a predefined list of possible words with theirrespective stem words is considered.
2.3.1. For stemming, a list of suffix words is stored into an array with their respective root words.
2.3.2. Consider check_token = availability in considered array of root word
2.3.3. If suffix of check_token = true, stem the word to its respective root word from the list of arrays.
2.3.4. Else, there is no need of stemming. Word is already in its root word format. Move to the next token.
Step 3Apply Correlation based feature selection approach to select the useful feature words from the pre-processed data. Correlation based feature selection method only selects the feature set which are most related to the particular class. If f is the feature set with k number of features and c is number of classes then CFS can be applied as mentioned in Equation 2.Where, rcf is the average of feature-class correlation. rcf is the average of feature-class correlation , rff is the average of feature-feature correlation.
.
Step 4: Calculate the probability distribution of the tokens along with selected features using NB approach. The formulation for the probability distribution is presented in Equation 3.Where, f is any feature vector set (f1, f2, f3,....fn) and y are the class variables with m possible outcomes (y1, y2, y3,....yn). P(y|x) stands for posterior probability, P(x|y) is any particular class on which P(y|x) is dependent. P(x) is evidence depending on the known feature variables, P(y) is the prior probability.
.
Step 5:Apply PSO approach to optimize the parameters of NB approach.
5.1. All the tokens are considered as the particles. Initially these particles randomly fly and search for the food sources in the form of the best feature match for tokens. Then search forthe local and global solution.
5.2. The performance of each particle depends upon the similarity with the features that hasto optimize.
5.3. Each particle flies over the n- dimensional outer search space and keep updating the following information:
Xi – current position of particle
Pi – the personal best position of particle x
Vi – the current velocity of particle
5.4. The velocity updates in PSO can be calculated using the formula given below by Equation (4)
Now, Vi is the new velocity. So, the position of the particle updates with the velocity as defined with Equation (5)
5.5. Update the positions for each particle and store the global best solutions.
*Step 6:*Based on the evaluated feature similarity using PSO, classification of tokens is declared as spam or non spam.
*Step 7:*Further, final classification is performed by evaluating the probability of spam or non-spamtokens in a sentence.
7.1. If the probability value of spam tokens is more, then email is considered as spam email.
7.2. Else, email is considered as non-spam email.
*Step 8:*Store the email as spam or non-spam and repeat the process for all the emails.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006209|

In the subsection below, we have shown the functionality of our proposed system in the following steps:
1. Register: The interface allows "User" or "Admin" to sign up by providing credentials to register with the application.
2. Login: "User" or "Admin" will log in with their accounts.
3. Random email input: The system takes the random emailas aninput.
*4. Recommended output:*Spam filtering techniques provides us with an appropriate output based on the input given in the mail as Spam or Ham.

V. FUTURE SCOPE

In our proposed system we have used the Naive Bayes algorithm. As future work, we can extend the functionalities by applying improved vectorization techniques such as TF IDF (term frequency-inverse document frequency) to remove common words to generate more accurate results. This will be helpful to enable sharing and improvement in the system. We can also use larger data sets. We can also improve the validation of data for our system. Try variations of Naive Bayes and other classification models.There is a wide scope of enhancement in our project. Following enhancements can be done: Filtering of spams can be done on the basis of its contents. The spam email classification is very important in classifying e-mails and to separate e-mails that are spam or non-spam. This method can be used by big organization to distinguish good mails that is only the mails they wish to receive.

VI. CONCLUSION

Through the Naive Bayes classifier, we were able to improve upon the baseline for spam filtering. Naive Bayes has a very fast processing speed and allows for a small training set, hence is suitable for real-time spam filtering. Previous research showed that most spam classifications failed when exposed to text modifications. By creating an addition to Naive Bayes, we were able to improve the multi-class prediction ability. We also found that our new addition helped improve ham classification due to the high recall and precision rates. We are able to classify mails as SPAM mails or genuine mails. Some mails might be considered into suspicious list. Na¨ive Bayes classifier also can get highest precision that give highest percentage spam message manage to block if the data_set collect from single-mail accounts. With so many people,enterprises and companies using mails as their daily part of their lifes , it is manually difficult to handle all possible mails as our projects deals with limited amount of corpus

VII. ACKNOWLEDGEMENT

We would like to thank our Project Guide Prof. M. K. Nivangune and Prof. S. N. Shelke for his ideas, active guidance and practical advice that has helped us tremendously in our project. We would like to express our gratitude to Prof. B.B. Gite, Head of the Computer Department, for providing us with a healthy environment and facilities in the department. His cooperation and encouragement have helped us a lot throughout this project. Lastly, we are grateful to the University for giving us a chance to work on this topic.

REFERENCES

1.A.A. Akinyelu, A.O. Adewumi, Classification of phishing email using randomforest machine learning technique, J. Appl. Math. 6 (2016). Article ID 425731,Retrieved on July 12, 2017 from.
2.S. Albelwi, A. Mahmood, A framework for designing the architectures of deepconvolutional neural networks, Entropy 19 (6) (2017) 242
3.Mathswork, Detector Performance Analysis Using ROC Curves–MATLAB&Simulink Example, Retrieved August 11, 2017 from, 2016,http://www.mathworks.com/help/phased/examples/detector-performance-analysis-using-roc-curves.html
4.G. He, Spam Detection, 1st ed. 2007.sharma, a. and jain, D. (2017). A survey on spam detection.
5. En.wikipedia.org. (2017). Spamming. [online] Available at: https://en.wikipedia.org/wiki/Spamming [Accessed 29 Aug. 2017].
6. bot2, V. (2017). Email Spam Filtering : A python implementation with scikit-learn. [online] Machine Learning in Action. Available at: https://appliedmachinelearning.wordpress.com/2017/01/23/email-spam-filter-python-scikit-learn/
7.Edstrom, Detecting Spam with Artificial Neural Networks, Retrieved on August10, 2018 from, 2016,http://homepages.cae.wisc.edu/~ece539/project/s16/Edstrom_rpt.pdf.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006209|

8. A. Bhowmick and S. Hazarika, "Machine learning for e-mail spamfiltering: review, techniques and trends," https://arxiv.org/abs/1606.0104,2016, accessed: 2017.
"An analyst review of hotmail anti-spam technology,"https://www.lifewire.com, Radicati Group, Inc, 2010, accessed:2017.
9. H. Tschabitscher, "How many emails are sent every day?"https://www.lifewire.com/how-many-emails-are-sent-every-day-117121,2017, accessed: 2017.
10. K. Tretyakov, "Machine learning techniques in spam filtering," inDataMining Problem-Oriented Seminar, 2004.
11.A. Aski and N. Sourti, "Proposed efficient algorithm to filter spamusing machine," inPacific Science Review A: Natural Science andEngineering, vol. 18, 2016, pp. 145–149.
12.J. Rao and D. Reiley, "The economics of spam," inJournal of EconomicPerspectives, vol. 26, no. 3, 2012.
13. "Abayesianapproachtofilteringjunke-mail,"https://robotics.stanford.edu/users/sahami/papers-dir/spam.pdf, StanfordUniversity, accessed: 2017, 2018.
14. Graham-Cumming, "Does bayesian poisoning exist?"https://www.virusbulletin.com/virusbulletin/2006/02/does-bayesian-poisoning-exist/, 2006.
15.B Biggio, P. Laskov, and B. Nelson, "Poisoning attacks against supportvector machines," inProc. of the 29.th International Conference onMachine Learning, 2012, pp. 1467–1474.
16.J. Eberhardt, "Bayesian spam detection," inUniversity of MinnesotaMorris Digital Well, vol. 2, no. 1, 2015.
17. K. Asa and L. Eikvil, "Text categorisation: a survey,"https://www.nr.no/ eikvil/tmsurvey.pdf, 1999.
18. M. Sprengers, "The effects of different bayesian poison methods on thequality of the bayesian spam filter," inB.S. thesis, RadboudUniversityNijmegen, Nijmegen, Netherlands, 2009.
19.S. Raschka, "Naive bayes and text classification i: introduction and the-ory," https://arxiv.org/abs/1410.5329, Cornell University Library, 2014,2015, 2017, accessed: 2017, 2018.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details