iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35992

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI June 2021- Available at www.ijraset.com

Spam E-Mail Filtering: A Review of Techniques

Rahul Anandpara[1], Dr. Mrs. Ranu R. Tuteja[2]
[1, 2]*M.E Computer Science and Engineering, Prof. Ram Meghe Institute of technology & Research Badnera*

Abstract: Every year, the number of uninvited email received by the common email user will increase dramatically. According to IDC, Spam has accounted for 38 percent of the 31 billion emails sent each day in North America in 2004, up from 24 percent in 2002. keeping pace with amount of spam is that the quantity of filtering solutions out there to assist eliminate it. This paper describes in detail how several of the most common spam filtering technologies work, how effective they are at stopping spam, their strengths and weaknesses, and techniques used by spammers to circumvent them.
Keywords: Spam; Spam filtering; Techniques; False positive;

I. INTRODUCTION

Electronic mail is employed daily by lots of folks to speak round the globe and may be a mission-critical application for several businesses. Over the last decade, uninvited bulk email has become a serious drawback for email users. an amazing quantity of spam is flowing into users' mailboxes daily. Not only is spam frustrating for most email users, it strains the IT infrastructure of organizations and costs businesses billions of dollars in lost productivity. The necessity of effective spam filters increases. In this paper, we presented our study on various problems associated with spam and spam filtering methods, techniques.

II. LITERATURE SURVEY

In the paper of Suryawanshi, Shubhangi & Goswami, Anurag & Patil, Pramod [1] different classification algorithms are compared. The experiments performed on email dataset from Kaggle and UCI repository. their results show that the performance of the classifier depends on the number of features and size of the dataset. In predefined norms, SVM is a better classifier. But after comparison, it is observed that Ensemble Classifier gives promising results than the other classifiers and the speed of the testing is also better. In their experiment, one of the validity threat is: testing was performed on email dataset without taking in the evolving patterns in the emails which may affect the accuracy of a classifier. The future work may incorporate ensemble classifier with Drift Detection technique to address the drift issue in the email spam filtering.

After a thorough analysis, the study of Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. [2] results in several different observations especially in the realm of Machine Learning based proposition. It is noted that high adoption of supervised approaches is quite obvious, the reason behind this turns out to be a better consistency in the performance of the model. It has also been highlighted that certain algorithms, such as SVM and Naïve Bayes are in high demand. they have also come to the conclusion that single-algorithm anti-spam systems are quite common thus the potentiality of research into hybrid and multi-algorithm systems are quite promising. Besides, research that focuses on email header features excluding the 'subject' field, URLs within the email body and sender domain information need to substantially increase. Another important area that needs increasing attention is the addressing of 'Concept Drift', which would definitely make a system to perform optimally under gradual modification in spamming techniques and motives. In addition, the current way of dealing spam emails of phishing nature is not the most efficient as described, thus requires a more innovative approach that will take into account the different angles of the problem.

Email spam has become one of the most demanding research topics due to increasing cybercrime and increasing spammers. Different authors have used different methods with testing on different datasets to detect email spam. With analysis from the results of existing techniques, K. Agarwal and T. Kumar [3], have used an integrated approach of NB and PSO for email spam detection. 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 the NB approach to improve the accuracy, search space, and classification process. Correlation-based feature selection (CFS) is used as a feature selection approach. Experimentation is performed on the Ling spam dataset with the use of evaluation parameters of precision, recall, f-measure, and accuracy. From the evaluated results, it can be declared that the proposed integrated concept outperformed in comparison with the individual NB approach. For future directions, the Naïve Bayes approach can also be integrated with any other swarm optimization-based concept like ant colony optimization, artificial bee colony optimization, firefly algorithm, etc. Also, the NB approach can also be changed with any other machine learning-based algorithm to further improve the results.

In the paper of Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora [4], they applied three different algorithms to detect spam emails, one being a new approach to spam detection. they adapted the spam filter to each user's preferences and predicted mail spam or not by text mining and text recognizing by OCR library TESSERACT. Although methods used by us have many advantages, it certainly does come with some disadvantages. The disadvantage of text filtering is that they are time-consuming. The OCR-based detection also has disadvantages like, the recognition is not always perfect, and works for certain fonts only, cannot predict for CAPTCHA images, and obviously are expensive. In the future, as it is an adaptable and scalable project thus they would like to detect threats found in emails that are viruses.

Feature selection is one of the important tasks in the classification process. The paper of Mohamad, Masurah, and Ali Selamat[5] has presented a hybrid feature selection method, namely the Hybrid feature Selection, which integrates the term-document frequency (TF-IDF) and the rough set theory to increase the classification result, generally and email filtering, specifically.

The Spam is a standout amongst the most irritating and malicious increments to the worldwide PC world. In the paper of Shradhanjali, Prof. Toran Verma[6], proposes a novel method for email spam detection which can effectively identify spam emails from their contents. The spam emails can be blocked by the user and genuine mail can be retained by the user. The proposed classifier achieves 98 % accuracy whiles classifying the series of datasets.

In the paper of W.A, Awad & S.M, ELseuofi [7], they review some of the most popular machine learning methods and of their applicability to the problem of spam e-mail classification. Descriptions of the algorithms are presented, and the comparison of their performance on the SpamAssassin spam corpus is presented, the experiment showing very promising results especially in the algorithms that are not popular in the commercial e-mail filtering packages, spam recall percentage in the six methods has the less value among the precision and the accuracy values, while in term of accuracy we can find that the Naïve Bayes and rough sets methods have a very satisfying performance among the other methods, more research has to be done to escalate the performance of the Naïve Bayes and Artificial immune system either by a hybrid system or by resolving the feature dependence issue in the naïve Bayes classifier, or hybrid the Immune by rough sets. Finally, hybrid systems look to be the most efficient way to generate a successful anti-spam filter nowadays. In the paper of A. K. Ameen and B. Kaya[8], they selected deep learning referred to spam detection method in Twitter. For this purpose, apart from the classical approaches, they first used features extracted in the content of tweets by the Word2Vec method. Then, they employed MLP neural network as the classification method. Finally, they compared the selected method with three different classifiers. The experiments show that the selected approach shows the best results in terms of precision, recall, and F-measure. In the paper of Tasnim Kabir, Abida Sanjana Shemonti, Atif Hasan Rahman[10], provide an efficient approach to the classification problem of assigning an unknown specimen to a known species by utilizing its DNA barcode. they implement this task using different supervised machine learning classification algorithms implemented using the software tool Weka. they use function-based method simple logistic function, lazy classifier instance-based k-nearest neighbor, tree-based classifier random forest, rule-based classifier PART, meta based attribute selected classifier and bagging and they test them on synthetic and empirical datasets belonging to animals, plants, and fungus kingdoms. In their work, they show with experiments that, part of the DNA sequence can be used to identify specimens effectively and efficiently, with the help of machine learning approaches. they compare the classification results with well-established DNA barcode classification techniques, such as phylogenetic trees, similarity-based BLAST, and character-based DNA-BAR, BLOG methods. From these comparisons, they can conclude that supervised machine learning algorithms can be explored as a promising technique for handling specimen classification by using DNA barcode sequences. These algorithms show excellent performance in accuracy, recall, and precision. The results of their work and the state-of-art methods establish the validity of supervised learning methods for species identification with DNA Barcode sequences. they test the performance of different supervised learning methods on different datasets revealing effective methods for species identification and classification with DNA Barcode sequences and show that their implemented algorithms show 6% improvement on average, compared to the state-of-art classification methods.

III.SPAM E-MAIL FILTERING TECHNIQUES

A. Signature Matching

One of the identifying characteristics of spam is that there's a flood of it (most definitions of spam deliberately embody the word "bulk"). Spammers send a duplicate of their spam message to each valid email account they'll realize. Signature matching takes advantage of this by mechanically discarding every copy of a spam message as before long because it acknowledges it as spam.

Vendors of signature matching anti-spam software package maintain an oversized variety of take a look at accounts at ISPs and free email services like Hotmail and Yahoo. They monitor these accounts closely, anticipating a spam message to arrive. once a spam message will arrive, the seller quickly generates a signature for that message.

Sometimes the signature could be a string of thirty two to 128 alphamerical digits that's calculated supported the content of the message. This signature is else to a info of all of the spam signatures that the seller has calculated.
Sites mistreatment the signature matching software package ar given a duplicate of this info by the antispam software package merchant. This info is put in on their mail server, and is updated on a really frequent basis. once the location receives a message, it generates a signature for it mistreatment precisely the same technique that their anti-spam merchant uses. to see if the message is spam, the anti-spam software merely checks to visualize if the signature for the incoming message matches any of the signatures in the spam signature info. If it does, then the message is treated as spam.

B. Heuristics

One of the characteristic characteristics of spam is that there's a flood of it (most definitions of spam deliberately embody the word "bulk"). Spammers send a duplicate of their spam message to each valid email account they'll notice. Signature matching takes advantage of this by mechanically discarding every copy of a spam message as shortly because it acknowledges it as spam.
Vendors of signature matching anti-spam code maintain an outsized range of take a look at accounts at ISPs and free email services like Hotmail and Yahoo. They monitor these accounts closely, anticipating a spam message to arrive. once a spam message will arrive, the seller quickly generates a signature for that message. sometimes the signature may be a string of thirty two to 128 character set digits that's calculated supported the content of the message. This signature is value-added to a information of all of the spam signatures that the seller has calculated.
Sites victimization the signature matching code ar supplied with a duplicate of this information by the antispam code marketer. This information is put in on their mail server, and is updated on a awfully frequent basis. once the positioning receives a message, it generates a signature for it victimization precisely the same technique that their anti-spam marketer uses. to see if the message is spam, the anti-spam software merely checks to check if the signature for the incoming message matches any of the signatures in the spam signature information. If it does, then the message is treated as spam.

C. Bayesian Filtering

Although they need been used for years to perform text classification, Bayesian filters area unit one in all the newest technologies used for filtering spam. The filters "learn" the distinction between spam and non-spam messages, and that they unceasingly update their data to remain current with new spam messages.
A Bayesian filter is educated the distinction between spam and non-spam mail by watching 2 massive collections of email messages. One assortment contains spam messages received by a website, and the other assortment contains non-spam messages received by identical website. In essence, the filter picks each message apart into individual words. supported a comparison of however typically a given word seems in spam messages as hostile non-spam messages, the filter calculates the chance that a message containing that given word is spam.
When a brand new message is received by the filter, it's force apart into individual words. The Bayesian filter chooses the words from the message that it thinks area unit the foremost attention-grabbing. probabilities of every of these words showing in a very spam message area unit combined victimization Bayes' Formula, and also the result's wont to verify if the message is spam.

D. DNS Blacklisting

One of the oldest sorts of spam hindrance, DNS blacklisting uses a centralized info to dam all email from a number getting used to send spam. The supplier of the blacklisting service maintains the database, adding entries for hosts that area unit getting used by spammers. Access to many of those databases is free, whereas others need a yearly fee for usage. During associate SMTP dealings, associate email server organized to use a DNS blacklist can perform a DNS query on the host that's causing the message. instead of acting the question against its own DNS server, the e-mail server queries a DNS server provided by the DNS blacklisting service. supported the information came from the question, the e-mail server can either settle for or reject the incoming message.

E. Challenge/Response

Virtually each spam message is generated and sent by an automatic computer code utility (spammers don't sit before of a laptop in their basement clicking the "Send" button as quick as they can). Challenge/response systems profit of this by forcing email senders to prove that they're human through some form of check (the "challenge"). When Associate in Nursing email message is shipped to Associate in Nursing account protected by a challenge/response system, it's placed in a holding space and a message containing a challenge is shipped back to the sender. typically this challenge message contains a quick rationalization of why it absolutely was sent, and includes a link to an internet page where the particular challenge are conferred.

If the message sender passes the challenge, the first message is discharged from the holding space and sent to the supposed recipient. If the message sender doesn't pass the challenge, then the first message is deleted once a fixed amount of your time. For a challenge to be effective it's to be one thing that humans will do simply however computers cannot. The most common style of challenge consists of a picture of distorted text. To pass the challenge, a human should sort the text properly.

IV. CONCLUSIONS

In this paper, we analyze spam e-mail filtering techniques that are commonly available today, and several more are under development. No single filtering method is a panacea for the spam problem, since each has weaknesses that spammers can exploit. The best answer is to use totally different, overlapping ways in parallel with each other. whereas a transmitter could also be able to craft messages that may sneak by one type of filter, it's just about not possible to put in writing a message that may evade multiple filtering ways. At an equivalent time, it's necessary to not use too several filtering ways at an equivalent time. Each one has a noticeable effect on email server performance. After messages have passed through two or three filtering methods, the additional accuracy imparted by additional methods is going to be minimal.

V. ACKNOWLEDGMENT

It is our proud privilege to release the feelings of our gratitude to every person who helped us directly or indirectly to conduct this research work. we express our heart full indebtness and owe a deep sense of gratitude to Prof. Sangram Dande for their sincere guidance and inspiration for completing this paper.

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 Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295. [08907831]. https://doi.org/10.1109/ACCESS.2019.2954791

[3] K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.

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

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

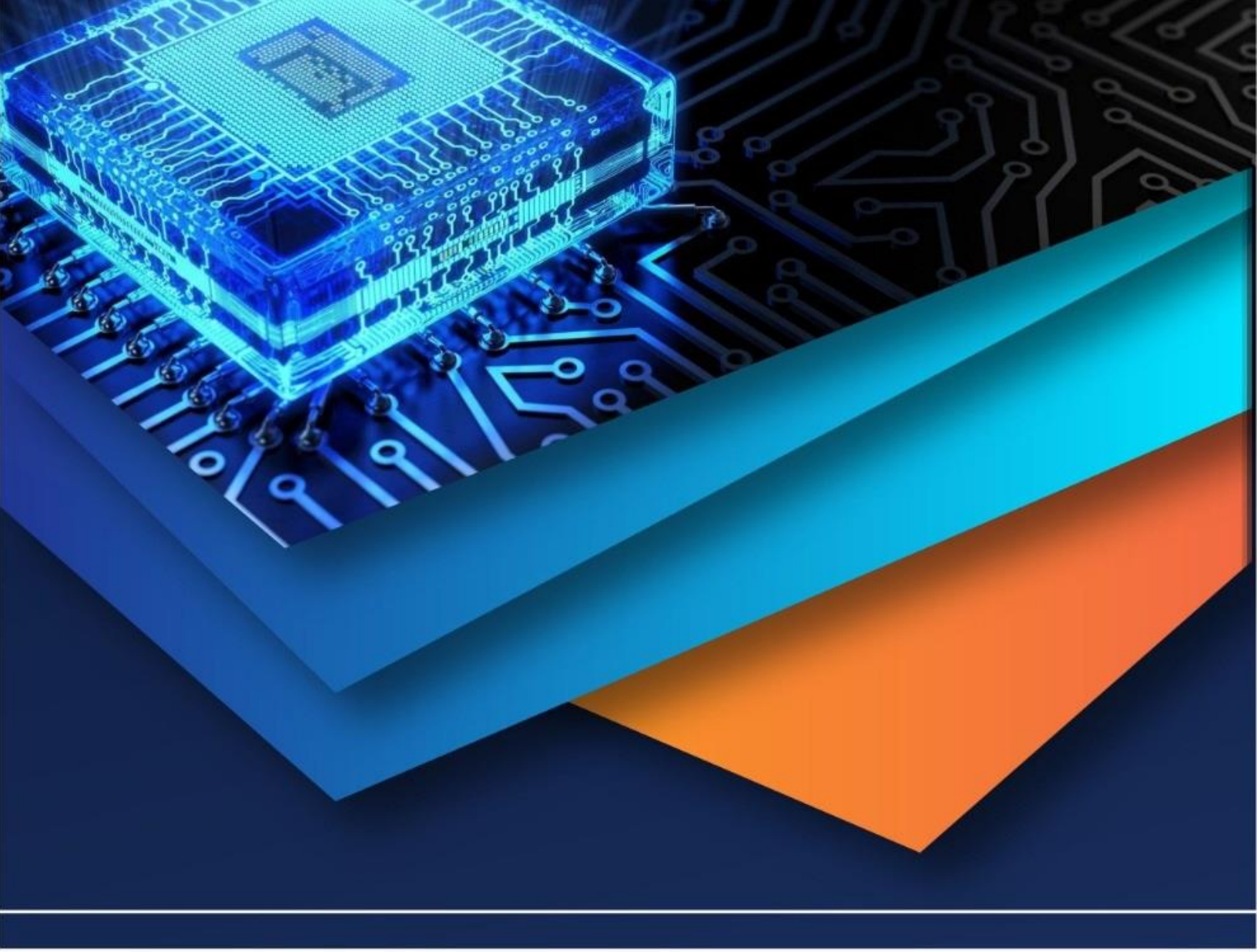
[6] Shradhanjali, Prof. Toran Verma "E-Mail Spam Detection and Classification Using SVM and Feature Extraction"in International Jouranl Of Advance Reasearch, Ideas and Innovation In Technology,2017 ISSN: 2454-132X Impact factor: 4.295

[7] W.A, Awad & S.M, ELseuofi. (2011). Machine Learning Methods for Spam E-Mail Classification. International Journal of Computer Science & Information Technology. 3. 10.5121/ijcsit.2011.3112.

[8] A. K. Ameen and B. Kaya, "Spam detection in online social networks by deep learning," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-4.

[9] Diren, D.D., Boran, S., Selvi, I.H., & Hatipoglu, T. (2019). Root Cause Detection with an Ensemble Machine Learning Approach in the Multivariate Manufacturing Process.

[10] Tasnim Kabir, Abida Sanjana Shemonti, Atif Hasan Rahman. "Notice of Violation of IEEE Publication Principles: Species Identification Using Partial DNA Sequence: A Machine Learning Approach", 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY