International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

SPAM AND EMAIL NOTIFICATION USING ARTIFICIAL NEURAL NETWORK ALGORITHM

M.Geethapriya[1], M.Jeevanantham[2], S.Raampradaap[3] ,A.Saravanakumar[4]

[1]Assistant Professor, Dept of Computer Science Engineering
[2,3,4] Dept of Computer Science Engineering
[1,2,3,4]N.S.N College of Engineering and Technology , Karur, India

---

Abstract -

Email spam is operations which are sending the undesirable messages to different email client. E-mail spam is the very recent problem for every individual. The e-mail spam is nothing it's an advertisement of any company/product or any kind of virus which is receiving by the email client mailbox without any notification. The detection of email spam involves various techniques such as rule-based filtering, content-based filtering, and machine learning algorithms. Rule-based filtering uses predefined rules to identify spam based on certain characteristics such as the sender's address, subject line, or message content. Content-based filtering analyses the content of the email, including keywords, images, and formatting, to identify spam. To solve this problem the different spam filtering technique is used. The spam filtering techniques are used to protect our mailbox for spam mails. The Artificial neural network classification with three-layer framework are used that includes obfuscator, classifier and anomaly detector for spam classification for bulk emails. The ANN is very simple and efficient method for spam classification. The real time dataset is used for classification of spam and non-spam mails. The feature extraction technique is used to extract the feature in terms of digest based on bucket classification. The result is to increase the accuracy of the system. And implement Self Acknowledgeable Intranet Mail System has been designed and implemented to benefit the sender about the status of his mail. Once a mail is sent, the sender can know the receiver activity in the mail system until the mail is viewed. Finally provide the pop-up window to identify the mail content at the time of open the spam mails.

Key Words: Big Data, Artificial Neural Network, .NET, SQL, Spam Mail

1.INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing application software's are inadequate to deal with them. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem.

Analysis of data sets can find new correlations to spot business trends, prevent diseases, and combat crime and so on. Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, finance, urban informatics, and business informatics.

Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled 2 every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte's (2.5×1018) of data are

generated Relational database management systems and desktop statistics- and visualization-packages often have difficulty handling big data. The work may require massively parallel software running on tens, hundreds, or even thousands of servers. What counts as big data varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration

2.LITERATURE REVIEW

In the modern era, all services are maintained online and everyone use it to speed up their day-to-day activities. This includes social as well as financial activities which involves usage of sensitive information to carry out the intended task. With the increase in usage of such facilities put forth the importance of securing the data used to perform such actions. Over the last decade phishing has become a serious threat to the society by stealing sensitive information to get hold of these facilities. This is considered to be the most profitable cybercrime and according to IBMs X-Force researcher's statistics, the number of people becoming the victim of such activities are increasing tremendously. As the risk of phishing emails are increasing steadily, the need to detect and overcome such situations stands as one of the highest priority tasks at hand. One of the drawbacks with the current model is that the proposed mechanism relies on feature selection, which requires domain knowledge. To overcome this issue deep learning models can be incorporated, which can learn more complex patterns from the raw data and use it as features that produce more efficacy and this can be considered as a possible future work. In addition to that both the subtasks belong to unconstrained category, allowing external datasets to be used for the training purpose. The datasets provided in the subtasks are highly imbalanced.

3.EXISTING SYSTEM

Many spam filtering techniques have been put into business, which include Bayesian spam filtering and collaborative filtering. The concept of Bayesian spam filters is proved to be remarkably efficient. However, it is difficult to detect all spam as spammers present many challenges to this content-based filtering technique, like changing vocabulary, introducing the most recognizable terms or adding a relatively high number of random words. The process of spam detection is similar to how memory is developed in our brain, as our spam detecting system can distinguish spam from non-spam emails based on a self-learning algorithm according to the principles of memory forming. The arrival of the new email can be treated as the excitatory input to each existing item, and the scale of the input is analogous to the similarity between the new email and each existing email item in the database. The strength of each item is then accumulated, i.e., the more the item resembles the new email, the stronger the stimulation is, and the faster the corresponding strength grows. When the strength value of an item exceeds the 'remembered threshold', it will be defined as 'spam' by the system.

On the contrary, while there is no more newly entering similar emails, the strength value of the corresponding item will decrease. When it drops below an inhibitory threshold, called the 'forgotten threshold', the item is deleted from the database. A sequence of chunk hashes is created to represent the text. Each text file is firstly partitioned into a sequence of chunks according to the algorithm TTTD (Two Thresholds, Two Divisors). The chunks of suspicious emails are then encoded by hash function, which is able to provide privacy for email users

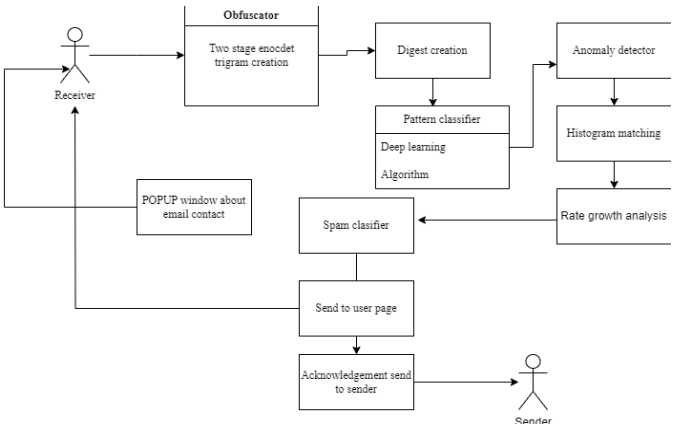
4.PROPOSED SYSTEM

Email is one of the crucial aspects of web data communication. The increasing use of email has led to a lucrative business opportunity called spamming. A spam is an unwanted data that a web user receives in the form of email or messages. This spamming is actually done by sending unsolicited bulk messages to indiscriminate set of recipients for advertising purpose. These spams messages not only increase the network communication and memory space but can also be used for some attack. This attack can be used to destroy user's information or reveal his identity or data. Spam emails are the emails that the receiver does not wish to receive. A large number of identical messages are sent to several recipients of email. Increasing volume of such spam emails is causing serious problems for internet users, Internet Service Providers, and the whole Internet backbone network. This may be denial of service where spammers send a huge traffic to an email server thus delaying legitimate message to reach intended recipients. Spam emails not only waste resources such as bandwidth, storage and computation power, but may contain fraudulent schemes, bogus offers and scheme. Apart from this, the time and energy of email receivers is wasted who must search for legitimate emails among the spam and take action to dispose the spam.

Dealing with spam and classifying it is a very difficult task. Moreover, a single model cannot tackle the problem since new spams are constantly evolving and these spams are often actively tailored so that they are not detected adding further impediment to accurate detection. Spamdoop is a platform that allows multiple entities to collaborate in early detection of bulk spam campaigns. In this project implement the Obfuscator which is used to encode the email content. And implement parallel classifier and propose anomaly detection approach to analyze the spam and normal mails. Finally provide email acknowledgement system to identify the view status of recipients with pop-up windows for email content.

5.SYSTEM ARCHITECTURE

Artificial neural networks (ANNs) are a class of machine learning algorithms that have been used for email spam detection. ANNs are particularly useful for classification problems, such as determining whether an email is spam or non-spam, as they can learn to recognize patterns in data and make accurate predictions. The process of using an ANN for email spam detection typically involves several steps. First, a dataset of known spam and non-spam emails is collected and used to train the ANN. The ANN is then tested on a separate dataset of emails to evaluate its accuracy and identify any areas where it may be misclassifying emails.



Architecture Description: One way to improve the accuracy of email spam detection using ANNs is to incorporate an acknowledgement mechanism. This mechanism involves sending an acknowledgement email to the sender after an email has been classified as spam. By incorporating an acknowledgement mechanism, the ANN can improve its accuracy over time. When a sender receives an acknowledgement email, they can modify their email content or behaviour to avoid having their emails flagged as spam in the future. This feedback can be used to update the ANN and improve its ability to accurately classify emails as spam or non-spam.

MAIL SERVER FRAMEWORK

To detect unsolicited and unwanted email and prevent those unwanted messages from getting to user's inbox is called spam filter. A mail server (also known as a mail transfer agent or MTA, a mail transport agent, a mail router or an Internet mailer) is an application that receives incoming e-mail from local users (people within the same domain) and remote senders and forwards outgoing e-mail for delivery. A computer dedicated to running such applications is also called a mail server. In this module we can create the framework like as mail server. This framework contains server and multiple users. Server can maintain all user details. Users easily upload the files in inbox and also share the data anywhere and anytime.

OBFUSCATOR

Obfuscation is the obscuring of the intended meaning of communication by making the message difficult to understand, usually with confusing and ambiguous language. The obfuscation might be either unintentional or intentional (although intent usually is connoted), and is accomplished with circumlocution, the use of jargon, and the use of an argot (in group language) of limited communicative value to outsiders. In this module, we can encode the emails based on ANN algorithm. The obfuscator can be customized to fit the organizational needs where administrators can decide their preferred cryptographic hashing function. It includes two stage encoding algorithm in Map Reduce platform. E-mails received by a participant are written into an index for batch processing. The body of the collected e-mails listed in the batch index is extracted

PARALLEL CLASSIFIER

In this module routes related digests to the same bucket and groups data so that it is stored on the same node. The routing of related digests is performed using a simple and efficient method that is applied. Using hashes of IP addresses to route Netflow information to the same nodes based on Convolutional neural network Classifier. In this module, we first train them first so it can build them up. After training the word probabilities are used to compute the probability that an email having particular set of words in it belong to either spam or legitimate emails. Each particular word or only the most interesting words contribute to email's spam probability. This contribution is known as the posterior probability and is computed using Bayes' theorem. Then, the emails spam probability is computed all over the word in the emails.

Using an Artificial Neural Network (ANN) algorithm for email spam detection involves several steps. The first step is to collect a dataset of known spam and non-spam emails, which will be used to train the ANN. This dataset should be diverse and include a variety of different types of spam emails. The second step is to preprocess the dataset, which involves cleaning and preparing the emails for input into the ANN. This can include tasks such as removing stop words, tokenizing the text, and converting the text into a numerical representation that can be used by the ANN. Next, the ANN architecture must be defined. This involves selecting the number of input and output neurons, as well as the number of hidden layers and neurons in each layer. The activation function for each neuron must also be selected, which determines how the neuron will respond to input.

Once the ANN architecture is defined, the next step is to train the ANN using the preprocessed dataset. This involves presenting the ANN with the input (preprocessed emails) and the output (spam or non-spam label) and adjusting the weights and biases of the neurons to minimize the error between the predicted output and the actual output. After the ANN is trained, the next step is to evaluate its performance on a separate dataset of emails that were not used in training. This evaluation step helps to determine the accuracy of the ANN and identify any areas where it may be misclassifying emails. Finally, the ANN can be used for email spam detection in real-time. Emails are preprocessed and input into the ANN, which predicts whether the email is spam or non-spam. Depending on the application, an acknowledgement or feedback mechanism may be incorporated to improve the accuracy of the ANN over time.

ANOMALY DETECTOR

In this module, histogram-based anomaly detection technique that has been used successfully for many other applications, including finding outlying instances in network traffic or system calls in computers indicating compromised systems. Finally constructing the e-mail occurrence density function; for each number of occurrences, we compute the number of messages that are present in the batch that many number of times. A well-known trick for the spammer to avoid detection is to introduce variations in the messages in order to decrease the occurrence and rate of any individual version. However, our implementation computes the score based on the density function of our proposed encoding of messages, which brings multiple versions of the message under the same digest. If the rate growth is high means automatically consider as spam emails.

POP-UP WINDOW

A window that suddenly appears (pops up) when you select an option with a mouse or press a special function key. Usually, the pop-up window contains a menu of commands and stays on the screen only until you select one of the commands. It then disappears. The user can easily open the mail contents, so it can easily affect the systems. So, we can design the pop-up window, it will give the details of overall mail contents as alert system.

MAIL ACKNOWLEDGEMENT

In this module, requesting a receipt does not guarantee that you will get one, for several reasons. Not all email applications or services support read receipts, and users can generally disable the functionality if they so wish. Those that do support it are not necessarily compatible with or capable of recognizing requests from a different email service or application. Generally, read receipts are only useful within an organization where all employees/members are using the same email service and application. In this module, send the acknowledgement to send about the status of email based on receiver. Email tracking is used by individuals, email marketers, spammers and phishers, to verify that emails are actually read by recipients, that email addresses are valid. Service Module handling the regular/periodic updating of user location in remote server.

6. CONCLUSIONS

E-mail is an efficient, quick and low-cost communication approach. E-mail Spam is non-requested data sent to the E-mail boxes. Spam could be a huge drawback each for users and for ISPs. According to investigation nowadays user receives a lot of spam emails then non spam emails. To avoid spam/irrelevant mails like effective spam filtering strategies. Spam mails area unit used for spreading virus or malicious code, for fraud in banking, for phishing, and for advertising. Spam messages are nuisance and huge problem to most users since they clutter their mailboxes and waste their time to delete all the junk mails before reading the legitimate ones. They also cost user money with dial up connections; waste network bandwidth and disk space. Bayesian classifier is one of the most important and widely used classifier and also, it's the simplest classification method due to its manipulating capabilities of tokens and associated probabilities according to the users' classification decision and empirical performance. In this project, to implemented the system to analyze each and every mail. And also provide privacy-based detection system to encode the emails using Digest based system. Enhance the anomaly detector, to predict emails with pop up window with email tracking system. In the future work we have a plan to implement other algorithm to our classification method to achieve better performance.

REFERENCES

[7] Salloum, Said, et al. "Phishing email detection using natural language processing techniques: a literature survey." Procedia Computer Science 189 (2021): 19-28.

[5] Zamir, Ammara, et al. "Phishing web site detection using diverse machine learning algorithms." The Electronic Library (2020).

[8] Bibi, Asma, et al. "Spam mail scanning using machine learning algorithm." J. Comput. 15.2 (2020): 73-84.

[1] Karim, Asif, et al. "A comprehensive survey for intelligent spam email detection." IEEE Access 7 (2019): 168261-168295.

[10] Fang, Yong, et al. "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism." IEEE Access 7 (2019): 56329-56340.

[2] Sharma, Priti, and Uma Bhardwaj. "Machine learning based spam e-mail detection." International Journal of Intelligent Engineering and Systems 11.3 (2018): 1-10.

[3] Vazhayil, Anu, et al. "PED-ML: Phishing email detection using classical machine learning techniques." Proc. 1st antiphishing shared pilot 4th acm int. workshop secur. privacy anal.(iwspa). Tempe, AZ, USA, 2018.

[6] Gupta, Mehul, et al. "A comparative study of spam SMS detection using machine learning classifiers." 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, 2018.

[9] Bhuiyan, Hanif, et al. "A survey of existing e-mail spam filtering methods considering machine learning

techniques." Global Journal of Computer Science and Technology (2018).

[4] Harikrishnan, N. B., R. Vinayakumar, and K. P. Soman. "A machine learning approach towards phishing email detection." Proceedings of the Anti-Phishing Pilot at ACM International Workshop on Security and Privacy Analytics (IWSPA AP). Vol. 2013. 2018.