A Review of Phishing Email Detection based on Different Machine Learning Methods

Sharon Abraham
M.Tech student, CSE Dept.
Mangalam College of Engineering,
Ettumannor,Kottayam India

Dr. Sabu George
Associate .Professor, CSE Dept.
Mangalam College of Engineering,
Ettumannor,Kottayam India

***Abstract*:- Emails are considered as a method of communication in both personal and professional life. Sensitive and private information, such as banking details, credit reports, login information, etc., is frequently sent by email. Because of this, they are important to cybercriminals who might misuse the data. Phishing is a technique used by fraudsters to trick people into giving up sensitive information by seeming to come from reliable sources. In a phished email, the sender can trick you into giving up personal information. To identify whether a email received is phished various machine learning techniques can be used. In this paper, various detection techniques are compared. Based on the techinques used it can be classified as phished or not.**

Keywords: Phished email, Neural Network, SVM, Legitimate email.

1. INTRODUCTION

The rapid development of Internet technologies has immensely changed on-line users' experience, while security issues are also getting more overwhelming. Presently, new threats have the potential to seriously harm customers' machines as well as steal their money and personal information. Phishing is a serious concern among them and is a criminal activity that uses social engineering and technology to steal a victim's account details and identification data. The number of phishing detections increased by 46% in the first quarter of 2018 compared to the fourth quarter of 2017, according to a report from the Anti-Phishing Working Group (APWG) [1].According to the data, it could be clearly understood that phishing has shown an apparent upward trend in recent years. Similarly, the harm caused by phishing can be imagined as well. Phishing emails are categorized as spam messages. Users receive emails alleging to be from a legitimate company or bank and asking the user to follow an embedded link. The link will redirect the user to a fake website that requests confidential information, such as usernames, passwords or credit card numbers.

Detection of phishing emails has received a lot of attention recently due to their impact on users' security. Therefore, many techniques have been developed to detect phishing emails varying from communication-oriented techniques, such as authentication protocols, blacklisting, and white-listing, to content-based filtering techniques. The blacklisting and white-listing techniques have not proven though to be sufficiently efficient when used in different domains, and thus they are not commonly used. Meanwhile, the content-based phishing filters have been widely used and have proven to be of high efficiency. In light of this, researches have focused on content-based mechanism and on developing machine learning and data mining techniques based on the header and body of emails.

Phishing is a lucrative sort of fraud when the perpetrator fools the recipients and acquires private information from them. Users of phished emails may be instructed to open an attachment or click on a link to a website where they must enter sensitive data like passwords and credit card numbers. The phisher sends the messages to thousands of people, and while typically only a tiny proportion of receivers fall for the scam, it can yield significant financial rewards for the sender.

The risk of losing sensitive information to fraudsters has increased along with the continuous rise of technology and email use. In this study, machine learning techniques are used to identify phished emails. Machine Learning is a field of artificial intelligence in which the system is given the ability to learn without being explicitly programmed. Algorithms for supervised machine learning are utilised for classification in our model. Based on the known instances, supervised learning systems forecast the nature of unknown data. These algorithms are a subset of those used in machine learning, which learn from data repeatedly.

With the high usage of emails and growth in technologies, risk of losing valuable information to fraudsters has also been increasing. This paper focuses on comparing different machine learning algorithms used in the field for phishing email detection. The below sections include the challenges faced, various phishing email detection methods considered and compared. And from comparison each method accuracy and conclusion is made.

2. CHALLENGES

Phishing is a technique used to steel personal information for the purposes of identity theft and using fake e-mail messages that appear to come from legitimate businesses. This is typically accomplished by sending emails that appear to be from reputable sources in order to access someone's private and sensitive information. Phishing emails are the fastest-rising type of internet fraud used to steal financial information from victims and commit identity theft. Responding to phishing emails by entering the desired financial or personal information into pop-up windows, websites, or emails puts the individual and their institutions at danger. With the massive work exists for phishing email detection task, there is no set of features that has been determined as the best to detected phishing. Moreover, the same nondeterministic scenario is applied for the underling classification algorithm. Finally, there is a need to keep on enhancing the accuracy of the detection techniques. Overall the challenges faced includes: How to determine the best set of features to be used with phishing detection. How to select the best classification algorithm to be used for phishing detection. How to enhance the performance of the best selected features and classifiers.

3. RELATED WORKS

For classifying phished emails, Andronicus et al. employed a random forest machine learning classifier. They sought to increase classification accuracy while reducing the amount of features needed. We provide a highly accurate content-based phishing detection method.

In [2], authors put forth a model based on information taken from email headers and HTML bodies that are then categorised using feed forward neural networks. The outcomes show a categorization accuracy of 98.72 percent.

In [3] method uses a dataset of more than 7000 emails and a variety of features. A 99.5 percent overall accuracy is attained.

Gilchan Park et al. sought to extract reliable traits to distinguish between genuine and phished emails. Between phishing emails and authentic emails, their closeness in sentence structure and the distinction in the subjects and objects of their target verbs are compared.

The various phishing methods are examined in "Email Phishing: An Open Threat to Everyone," along with advice on how users can keep themselves out of scammers' traps.

C. Emilin Shyni et al proposed A methodology that combines natural language processing, machine learning, and image processing. They employ a total of 61 characteristics. Using a multi-classifier, they were able to attain a classification accuracy of over 96 percent.

4. PHISHING EMAIL DETECTION

One way to discern between legitimate and phished email communications is to filter emails. This method employs either a learning-based filter that analyses a collection of labelled coaching data or previously collected messages with upright assessments or a phishing e-mail filter that examines and groups emails into their suitable groups. Examining each email separately for any unique words is another way to analyse e-mail messages. The body and header of emails are separated [5]. Email headers include a number of fields, including from, subject, to, and others [5]. The header lines include explicit routing information in addition to information about the message's subject, receiver, and sender. The body of the email follows the header lines and contributes to the message’s content. Nowadays email phishing has become a big threat to all, and is increasing day by day[4].Different machine learning algorithms used are discussed below:

Support Vector Machine

SVM is typically utilised for both classification and regression tasks. The SVM plots each piece of data as a point in an n-dimensional space (n is the feature number for each sample within the training set). The algorithm's goal is to find the best hyper-plane, which may be divided into two types. SVM classifies the nonlinearly separable data by transforming it into a higher-dimensional space with the use of a kernel function that contains a separating hyperspace. The SVM is very memory sensitive and challenging to comprehend[6].

In phished email detection, input is represented as a set of features for instance, presence or absence of certain word or sentences and output will be 1 or -1 which indicates whether the email is phished or not.

Logistic Regression

When one or more independent (or predictor) variables are present, the binary logistic model is used to calculate the likelihood of a binary response (features). It enables one to state that the existence of a risk factor raises the likelihood of a specific result by a given percentage.

Neural Network (NN)

The structure of the NN is formed by a set of interconnected identical units called neurons. Signals are sent from one neuron to another via these connections. Weights are also affixed to the interconnections to improve delivery between the neurons. The neurons are weak on

their own, but when they are linked together, they can perform intricate calculations. Connectivity plays a big part throughout the testing phase since the interconnection weights are adjusted during network training. The NN illustration is shown in Figure 1. The input layer, hidden layer, and output layer of the NN are shown in the figure. The network is referred to as feedforward as the interconnections do not skip or loop back to the rest of the neurons. The nonlinearity present within hidden neurons helps provide the NNs power. Furthermore, the network must include nonlinearity so that complex mapping can be learnt.

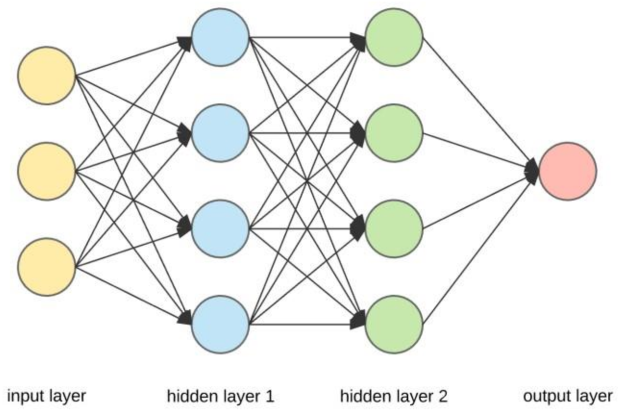

Fig:1 Neural Network

Random Forest

A random forest is an ensemble classifier that uses various decision trees to produce predictions. It operates by fitting various decision tree classifiers to various dataset subsamples. A random selection of the best qualities was also used to build each tree in the forest. Decision trees are generated (per the developer's specifications) during the training phase and used for class prediction[4]. They are obtained by taking into account the voted classes for each particular individual tree, with the class receiving the most votes being regarded the output.

As an ensemble learning technique for classification, regression, and other tasks, random forests or random decision forests build a large number of decision trees during the training phase and output the class that represents the mean of the classes (classification) or mean prediction (regression) of the individual trees. The tendency of decision trees to overfit their training set is corrected by random decision forests.

Naïve Bayes

This classifier uses the Bayes rule of conditional probability and applies to all data features. They are each examined separately under the presumption that they are equally essential to one another and independent of one another. Although the classifiers have the advantages of quick convergence and simplicity, it is impossible to comprehend the relationships and interactions between the attributes of each sample[7].

The naive bayes classifier, a member of the family of probabilistic algorithms, classified sample data using the Bayes theorem. theorem of Bayes According to Bayes' theorem, the probability of the hypothesis P(H) before receiving the evidence and the probability P(H|E) of the hypothesis following receipt of the evidence are related in the following ways:

$$P(H|E) =[P(E|H) / P(E)] *P(H)$$

Each category's probability is calculated, and the highest probability is the result.

5. COMPARISON

The different machine learning algorithms used were compared. The compared methods include SVM, Logistic regression, Neural Networks, Random Forest and Naïve Bayes. The comparison results are depicted in table 1.

Table 1 Comparison

Method	Precision	Recall	F measure	Accuracy
SVM	0.998	0.998	0.998	98.87
Logistic	0.956	0.956	0.956	95.63
Neural Network	0.999	0.999	0.999	99.87
Random Forest	0.999	0.999	0.999	99.87
Naïve Bayes	0.998	0.998	0.998	99.81

6. CONCLUSION

Phishing email is currently on the important topic in the field of cybersecurity. As the increase the number of phished cases the requirement for early detection has been a need. Hence different algorithms were introduced to this field to overcome the issues faced. Concerns about security issues have become more intense with developments in internet technologies and the consequent revolution in online user interaction.

Hence in this paper various techniques used in this field where compared. The compared machine learning techniques include SVM or Support Vector Machine, Neural Networks, Random Forest, Logistic Regression and Naïve Bayes. From the comparison it was observed that Neural network and Random Forest has higher detection accuracy than other methods.

REFERENCES

[1] A.-P. W. Group et al., “Phishing activity trends report 1st quarter 2018,”USA: Anti-Phishing Working Group (APWG), 2018.

[2] Andronicus A. Akinyelu and Aderemi O. Adewumi. Classification of Phishing Email using Random forest Machine Learning Technique 2014.

[3] Noor Ghazi M. Jameel, Loay E. George. Detection of Phishing Emails using Feed Forward Neural Network, International Journal of Computer Applications 2013.

[4] P. Verma, A. Goyal, and Y. Gigras, ‘‘Email phishing: Text classification using natural language processing,’’

[5] Comput. Sci. Inf. Technol., vol. 1, no. 1, pp. 1–12, May 2020, doi: 10.11591/csit.v1i1.p1-12.

[6] B. B. Gupta and Q. Z. Sheng, Machine Learning for Computer and Cyber Security: Principle, Algorithms, and Practices. Boca Raton, FL, USA: CRC Press, 2019.

[7] A. Kumar, J. M. Chatterjee, and V. G. Díaz, ‘‘A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing,’’ Int. J. Electr. Comput. Eng., vol. 10, no. 1, p. 486, Feb. 2020, doi: 10.11591/ijece.v10i1.pp486-493

[8] M. S. Swetha and G. Sarraf, ‘‘Spam email and malware elimination employing various classification techniques,’’ in Proc. 4th Int. Conf. Recent Trends Electron., Inf., Commun. Technol. (RTEICT), May 2019, pp. 140–145, doi: 10.1109/RTEICT46194.2019.9016964.