Phishing Website Detection Using Machine Learning: A Review

Yashaswini Girish SVCE Bengaluru India
Padmashree M, SVCE Bengaluru India
Vishaal Kumaar, SVCE Bengaluru India
N Jai Santosh SVCE Bengaluru India.
Dr Bama Hari SVCE Bengaluru India

ABSTRACT

Phishing is defined as fraudulent processes that aim to obtain some information, especially username, password, or any financial details, for access and malicious reasons. Phishing fraud could be one of the most common practices in the field of cybercrime to be amongst the one today. There are variety of domains where phishing attack can take place such as online payment sector, webmail, financial institution, file hosting or cloud storage and so on.
The webmail and online payment sector was more targeted by phishing than in any other industry sector.

There are several anti-phishing techniques that include blacklist, heuristic, visual similarity and machine learning. From this, blacklist approach is commonly using/used as it is very easy to use and implement, but it is not able to detect new phishing attacks. Machine Learning is an effective technique used to detect phishing. It also removes the drawback of the existing approach.

Our research includes a comprehensive literature review and proposes a new approach to detect phishing websites using feature extraction and machine learning algorithms.

Keywords: Phishing, Phishing Attack.

I. INTRODUCTION

Now a days, as there are so many people are being aware of using internet to perform various activities like online shopping, online bill payment, online mobile recharge, banking transaction. Due to wide use of this customer face various security threats like example of spam, fraud, cyber terrorisms and phishing. Of this, phishing is a new cybercrime and it is increasingly common nowadays. Phishing is an attempt of fraud, which has as the objective the theft of sensitive data such as username, passwords and credit card details. Information of user. Phisher design website looks like same as any legitimate site and spoof user for obtaining private information of user such as username, password, banking details for miscellaneous reasons.

According to APWG 2Q report [2]. The total number The instances in 2Q 2018 are significantly higher than the 13% it had in 1Q 2018. Its total reached 103,291 instances in 2Q 2018, compared to 70,611 in 1Q 2018.

There were increases in SAAS/webmail targeted sector, which had 21% of overall phishing attack. sector is proceeding as most attractive target for phishing.

According to APWG 1Q report [3], the total number of phish detected in 1Q 2018 was 263,538. That was up46 percent from the 180,577 observed in 4Q 2017.This was also well above the 190,942 seen in 3Q 2017. The number of unique phishing reports sent to APWG in 1Q 2018 was 262,704, compared to 233,613 in 4Q 2017 and 296,208 in 3Q 2017.

In the second quarter of 2023, the APWG observed 1,286,208 phishing attacks.
This was the third-highest quarterly total that the APWG has ever recorded.

II. BACKGROUND THEORY

Phishing involves the attempt to fraudulently acquire sensitive information, such as user names, passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication [1].Phishing can be done by different methods like email phishing, website phishing, spear phishing, whaling, tab napping, and evil twin type. There are different methods to stop such attacks, like the blacklists, heuristics, visual similarity checks, machine learning, and others.

A. Blacklist method

The most common technique used is the blacklist method, which entails that there is a database for the phishing URLs. If the URL is found in the database, it's flagged as a phishing URL and then informs the user for the same, else its legitimate. This method is straightforward and quick to deploy, as it simply checks if the URL is in the database. The major limitation of this technique is that a minor variation in the URL easily slips through detection using this list-based technique; so, regular updates of the list are critically important in keeping up with new attacks.

B. Heuristic based method

This extends the blacklist approach, using features extracted from phishing sites to detect a new phishing attack. This approach has its limitations though: it fails to detect all new attacks, and when an attacker understands the algorithm or features used, it becomes easier to bypass. Additionally, this method has poor detection rates because sites may or may not have common features.

C. Visual similarity

The visual similarity approach cheats on the users by a representation of an image of the actual site. However, this approach is also limited: the process is time-consuming and needs more space to compare two images of the sites, hence greater false negative rates. Additionally, it fails to detect when there are slight changes in the visual appearance of the site.

D. Machine learning

This method has been shown to be especially applicable to vast data and eliminates some difficulties experienced by other methodologies in use—for example, detecting zero-day attacks. By far, of all the techniques applied, the ones that have shown more efficiency through these years are the classifiers based on machine learning. The performance of these classifiers is based on a lot of factors like the size of the training data, the feature set used, and the type of classifier used. However, a limitation is that some of the approaches may fail to work in case attackers outsmart the filter through compromised domains from which to host their pages.

A lot of work has been done towards increasing accuracy in the detection of phishing websites by use of the myriad of classifiers. Some of the classifiers used are KNN, SVM, Decision Tree, ANN, Naïve Bayes, PART, ELM, and Random Forest. According to my study of the literature, the bigger the size of the dataset, the better the tree-based classifiers will operate, including Decision Trees and Random Forests. Therefore, the proposed approach for phishing website detection was to use tree- based classifiers.

Performance metrics for various algorithms include the F-measure, precision, recall, accuracy, AUC, ROC curve, among others.

III. LITERATURE SURVEY

Mohammad et al. [7] have proposed a model where important features that are essential to detect a phishing website are extracted automatically with minimal human intervention. Finally, the authors concluded that the process of their feature extraction is much faster andmore reliable than the manual process of extraction

Ahmad et al. [8] introduced three novel features in addition to the generally known features, to enhance the accuracy of classification of phished web pages. In this work, a set of new features and commonly known features for the classification of phishing and non-phishing sites are considered. They also concluded that their work could be further advanced through the introduction of such new features together with decision

tree machine learning classifiers.

In their approach, Pradeepthi et al. [9] researched other classification algorithms for detecting phishing URLs and found that the tree-based classifiers are the best among all that provide better accuracy. The authors have also used various features like lexical features, URL-based features, network-based features, and domain-based features.

Mustafa et al. [10] proposed a secure architecture for detecting phishing websites. The subset selection methods proposed are the extraction of features of websites using URL and then applying CFS subset-based and content subset selection techniques for the improvement of accuracy. Two subset selection methods are the CFS subset-based and content subset selection. The next stage for the purpose of classification is the use of machine learning algorithms.

Bhagyashree et al. [11] proposed the feature-based method to categorize URL as phishing or non-phishing. The method is based on the combination of a number of features, among which are lexical, WHOIS, Page Rank, Alexa rank, and Phish Tank-based features in categorizing an URL as phishing or not. The web-mining classification techniques to be used in their approach are based on these features.

Ahmad et al. [12] introduced three novel features in addition to the generally known features, to enhance the accuracy of classification of phished web pages. In this work, a set of new features and commonly known features for the classification of phishing and non-phishing sites are considered. They also concluded that their work could be further advanced through the introduction of such new features together with decision

A two-staged approach has been adopted by M. Amaad et al. [13] for the classification of phishing websites.
Their approach consists of first applying classification techniques independently, and the top three models are selected in accordance with the high accuracy and other criteria of performance. In the second phase, they ensemble each of the single models with the top three models to combine into a hybrid model that was above the accuracy achieved by the single model. They thus achieved an accuracy of 97.75% on the testing dataset.

One limitation of this model is that it requires more time

Priyanka et al. [12] proposed a novel approach of combining two or more algorithms. The authors implemented Adaline and Back propion along with Support Vector Machine (SVM) for achieving good detection rate and classification purposes.
The open-source framework, which was constructed by Hossein et al. [10] and was named "Fresh-Phish," is what one can use to obtain the data. Reduced feature set and Python are involved in the construction of queries. The framework enabled them to create a big labeled dataset, from which they tested a couple of machine learning classifiers. Analysis of these displayed very good accuracy. They have also done an analysis of the time it takes to train the models.

Gupta et al. [15] proposed a new anti-phishing technique based solely on the extraction of client-side features.
Here, the approach marked by its speed and reliability, meaning it is not dependent on a third party, but in contrast, extracts features solely from URLs and source code. They succeeded to have the overall detection accuracy of 99.09% while detecting phishing websites in their research. However, they have also claimed a limitation of their approach that it could only detect webpages written in HTML and couldn't detect non-HTML webpages.

Chunlin et al. [16] proposed an approach that mainly relies upon the character frequency features. They used statistical analysis of the URLs along with machine learning techniques to obtain improved accurate results in the classification of malicious URLs. Also, it was compared with the other six machine learning algorithms for the validation of its efficiency, where the proposed algorithm achieved a precision of 99.7% with afalse positive of 0.4%.

Sudhanshu et al. [17] used an association data mining technique for the identification of phishing websites. In the same regard, they proposed a special rule-based classification technique. They found that the association classification algorithm was better compared to others because of its simple rule transformation process. With 16 features, they managed to achieve 92.67% accuracy, leaving further room for the enhancement of the proposed algorithm to be more effective with achieving higher detection rates.

The open-source framework, which was constructed by Hossein et al. [14] and was named "Fresh-Phish," is what one can use to obtain the data. Reduced feature set and Python are involved in the construction of queries. The framework enabled them to create a big labeled dataset, from which they tested a couple of machine learning classifiers. Analysis of these displayed very good accuracy. They have also done an analysis of the time it takes to train the models.

CONCLUSION

Phishing is one of the ways in which to get hold of the user's personal information through deceitful emails or websites. Almost everything today is online, beginning from shopping for clothes to electronic gadgets, and paying off mobile, TV, or electricity bills. Now people are more tending online rather than standing in a line for hours. This shift has rather given phishers a huge scope to implement phishing scams. Although different approaches in the research areas have been developed for detection of phishing, yet no single approach could detect all kinds of phishing attacks. Since the technology is moving on, the phishing attackers use new methods; therefore, it is very important to find effective classifiers for the detection of phishing.

In this paper, we presented an overall literature survey on the detection of the phishing website. The machine learning approach tree-based classifiers, as per our findings, were best to suit these tasks as compared to the others.

I. REFERENCES

[1] Phishing definition, https://en.wikipedia.org/wiki/Phishing

[2] APWG Report1, http://docs.apwg.org/reports/apwg_trends_report_q2_2018.pdf

[3] APWG report2, http://docs.apwg.org/reports/apwg_trends_report_q1_2018.pdf

[4] Phishing dataset, https://www.phishtank.com/developer_info.php

[5] J. Han and M. Kamber, Data Mining Concepts and Techniques, Elsevier, 2006.

[6] Routhu Srinivasa Rao1 , Alwyn Roshan Pais : Detection of phishing websites using an efficient feature-based machine learning framework :In Springer 2018.

[7] Chunlin Liu, Bo Lang : Finding effective classifier for malicious URL detection : In ACM,2018

[8] Sudhanshu Gautam, Kritika Rani and Bansidhar Joshi : Detecting Phishing Websites Using Rule-Based Classification Algorithm: A Comparison : In Springer,2018.

[9] M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani : A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms :In International Conference on Computational Science and Computational Intelligence IEEE ,2016.

[10] Hossein Shirazi, Kyle Haefner, Indrakshi Ray: Fresh-Phish: A Framework for Auto-Detection of Phishing Websites: In (International Conference on Information Reuse andIntegration (IRI)) IEEE,2017.

[11] Ankit Kumar Jain, B. B. Gupta : Towards detection of phishing websites on client-sideusing machine learning based approach :In Springer Science+Business Media, LLC, part of Springer Nature 2017

[12] Bhagyashree E. Sananse, Tanuja K. Sarode :Phishing URL Detection: A Machine Learning and Web Mining-based Approach : In International Journal of Computer Applications,2015

[13] Mustafa AYDIN, Nazife BAYKAL : Feature Extraction and Classification Phishing Websites Based on URL : IEEE,2015

[14] Priyanka Singh, Yogendra P.S. Maravi, Sanjeev Sharma : Phishing Websites Detection through Supervised Learning Networks : In IEEE,2015

[15] Pradeepthi. K V and Kannan. A: Performance Study of Classification Techniques for Phishing URL Detection: In 2014 Sixth International Conference on Advanced Computing(ICoAC) IEEE,2014

[16] Luong Anh Tuan Nguyen†, Ba Lam To† ,Huu Khuong Nguyen† and Minh Hoang Nguyen : Detecting Phishing Web sites: A Heuristic URL-Based Approach: In The 2013 International Conference on Advanced Technologies for Communications (ATC'13)

[17] Ahmad Abunadi, Anazida Zainal ,Oluwatobi Akanb: Feature Extraction Process: A Phishing Detection Approach :In IEEE,2013.

[18] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey: An Assessment of Features Related to Phishing Websites using an Automated Technique:In The 7th International Conference for Internet Technology and Secured Transactions,IEEE,2012

[19] https://www.researchgate.net/publication/372056025_Phishing_Website_Detection_Using_Machine_Learning_A_Review