IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11552

Survey: Approaches for Phishing Detection

Abhishek Patil[1], Harshal Patil[2], Tejaswini Savkar[3], Priyanka Shirore[4], Prof D. M. Kanade[5]

Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik

Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik

Abstract: Internet has been a huge part of our day to day life. Since we are highly depended on Internet for all our daily activities, we are prone to cybercrimes. URL-based phishing attacks are one of the major threats facing by internet users. It is a way of fraudulent communication to steal the confidential data of user.Attackers mainly target people and reputed organizations, by tricking them to click on the URLs that seems to be secured and hence steal personal information of user or by injecting malware into machines.Researchers are constantly making several attempts to improve the accuracy and make model efficient.

In this paper, we aim to study and review various machine learning algorithms along with the datasets, that are used to detect legitimacy of the URL.The paper also provides statistical information about performance of the model. Our objective is to create a survey aid for researchers to examine the latest trends of phishing attacks and contribute in building phishing detection models that yield greater accuracy.

Index Terms: Phishing, Legitimate, URL features, machine learning, phishing detection

I INTRODUCTION

The year 2020 has seen enormous dependency of peo- ple on Internet due to Covid pandemic. All the physical work was shifted to virtual mode taking in consideration the importance of social distancing.Due to huge digital- ization, cyber criminals went on a web crime spree which further became a huge threat[1]. Generally in phishing attacks, attackers make use of websites to redirect users to sites,where they are fooled and forced into sharing user- names,passwords and other sensitive information such as bank details of the user. These phishing URLs can be sent to the target via email, instant message, or SMS. According to the 2020 FBI Criminal Registry, phishing attacks became the most common type of cyberattack in 2020, and phishing incidents almost doubled from 114,703 in 2019 to 241,343 in 2020 [2].

In online phishing, attackers tricks people into trusting
their websites, where they are tricked into sharing user- names, passwords, banking or credit card information, and other sensitive identifying information of User.This phishing URLs can be send to the people by email or SMS.According to an Atlas VPN investigation, Google detected a record-high number of phishing websites last year, reaching more than 2.11 million.In 2020, phishing sites jumped to 2.11 million, constituting a 25% growth over 2019, when the tech giant discovered 1.69 million malicious domains[2].

The number of phishing attacks detected using An- tiPhishing Work Group (APWG) increased in 2020, dou- bling during the year. Financial institutions are the most popular. In the 4th region of 2020, it turned into observed that phishing attacks in opposition to monetary establish- ments have been the maximum prevalent. As the most visited sites are E-trade,E-commerce sites,attacks against E-trade sites increased,while mail sites decreased[1].

Amid the triumph of the pandemic, there have been the most globally recognized phishing attacks on Covid19.According to WHO, many hackers and cybercriminals are sending emails and WhatsApp messages scam people, take advantage of corona virus disease [4] . Phishing is one of the mechanisms used by attackers to steal sensitive information needed for fraudulent transactions. According to Kaspersky Lab, their anti-phishing mechanism turned into prompted 246,231,645 times in KasperskyLab (2017)[5].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11552

Fig. 1. Scam Activity 2020

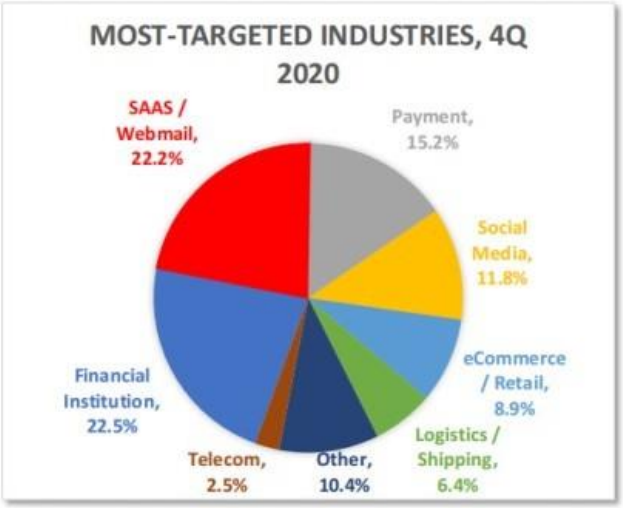

Fig. 2. Most Centric Industries, 4Q 2020

II BACKGROUND

A. Phishing Detection

Attackers carry out purely URL-based phishing attacks by sending malicious,fraudulent, hyperlinks that appear to be secured to the users and deceive them by forcing them to click on the links provided and hence stealing the sensitive information.In phishing detection mechanism, initially features are extracted for incoming URL, analyzed, various ML algorithms are applied and finally the URL is classified accordingly as phishing or legitimate.

B. Phishing Detection Approaches

There are many anti-phishing strategies and they are divided into 3 classes.

1) List-based techniques

In this case, a list is maintained such that valid websites or URLs are kept in the whitelist and phishing sites/URLs are kept in the blacklist [8].

These techniques make use of both whitelist and black-list based phishing detection techniques. List of suspicious URLs and IP addresses that are used to validate whether a URL is fraudulent are included in blacklist. Some of the prominent blacklists are managed using PhishTank and Google[6].

Blacklist and Whitelist are lists of websites that have been established as fraudulent and legitimate, respectively [7]. The given URLs in question may be matched with the two lists to decide if it's unsafe or safe. A given URL that doesn't exist withinside the white listing or explicitly exists withinside the black listing is probably a phishing website URL[8].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11552

DRAWBACKS:

There are also issues with the authenticity of these listings due to false positive and true negative assertion issues where a site can be valid and distinguished. Class is phishing or a website can be illegal and can be classified as valid. This method essentially turns the scenario into an endless race between security researchers and creator of phishing sites ,it is particularly tedious and difficult to do as each site wants to be checked very carefully before being declared a phishing site and listed. So new phishing sites frequently absent from this list, which can be dangerous for Internet users[9].

2) Visual similarity Based techniques

Visual similarity based techniques works through the use of visual similarity rating of web pages [10]. These statistics are then compared with image processing tech- niques.Evaluated web pages are converted into sentences of content or data and its attributes, where each web page is expected to incorporate three elements - Actual text, All visible elements, and overlook of the website.

The features of all three were extracted and used to generate an overall signature, and the signatures were compared using pairwise matching elements [12]. The right method here calculates the visual similarity score of each CSS selector used in the weighted site for each element of the site [11].

DRAWBACKS:

Although this approach is Language independent however require heavy computational power to process. Newly launched Website URLs are missing from the dataset listed. It Includes handling pre-processing of Inline or internal CSS CSS, causing a more overhead[7].

3) Content Based Approach

These techniques extract source code functionality from suspicious URLs.This approach makes use of features such as hyperlinks-based features, textual content-based fea- tures, tag-based features, image based features in various approaches collectively. Broken hyperlink,regular URL and empty URL are few of the Hyperlink-based features(Shirazi et al. 2018[13]; Rao and Pais 2018[14]; Marchal et al. 2017[15]; Jain and Gupta 2018 [16]). The text-based fea- tures are obtained by extracting essential keywords from the website.TF-IDF is one of the widely used approach for extracting important keywords

Several technical code content extraction techniques including names, titles, copyrights, anchor hyperlinks and domains are transmitted to third-party services to reduce false positive rates (FPR).Third-party services include use of search engines, site ratings, WHOIS, etc. Techniques in Jain and Gupta (2017) uses search engine results, Rao and Pais (2018) use site ratings and WHOIS to detect deceptive websites.

DRAWBACKS:

These strategies are computationally luxurious and require lots of statistics. They are complicated and conversation intensive. Content of legitimate websites may match with phishing websites' content material

4) Heuristic Based Approach

It is also known as a machine learning based phishing detection system. Phishing website detection in machine learning based system.This method uses classification of URL features using some artificial intelligence techniques. Generally URLs comprises of internet website features or website content,domain name, etc.

In the literature, there are a few works in this kind of detection mechanism. The formerly stated CANTINA project [17] achieved more accuracy with the help of the machine learning method approach. According to Tf-Idf and heuristic approaches, they detected a accuracy rate of 90% .

Researchers improved an anti-phishing security ma- chine which also called as "PhishWHO", implemented in

[18] ,in three steps to eliminate determining whether a website is legitimate or not with accuracy rate of 96.10%. In [19], phishing sites can be described using classifiers with URL attributes such as URL length, subdomain,delimiter domain and filename . Header and priority order

of incoming emails are covered in [20] .

In [21], URL-based features are used in conjunction with security-related features transport layer (length, num- ber of slashes, number of points and position). They discovered an accuracy rate of 93% with the help of using instructions received with the help of using apriori algorithm.

In [22], a non-linear regression method is used to decide whether or not a website is phishing. The

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11552

harmony search and Support Vector Machine (SVM) methods are used in the process. They used 11055 sites and 20 features. The envelope-priority decision tree approach is used for feature training in this method. They achieved accuracy of 92.80% .

In another study [23], a fraud detection system has been proposed, which includes 209 word vector features and 17 NLP-based features. The Random Forest, SMO, and Naive Bayes algorithms were compared, and the Random Forest algorithm in the hybrid method produced the best results in the proposed system, with an accuracy rate of 89.9%.

In [24], the diversity of NLP vectors was improved, and three different machine learning algorithms were compared based on their degree of precision .The Random Forest, SMO and Naive Bayes algorithms were compared. Random Forest hybrid method gave the best result with an accuracy rate of 97.2% .

The researchers created a phishing detection system in [25] with the use of neural network adaptive self- configuration for classification. In another study, 17 differ- ent features were used, which are widely used third-party services. Therefore, it was stated that much more time is needed in real-time implementation of study.

In [26], to distinguish phishing sites from legitimate ones, a utility learning method which is independent of any third parties, was used with 19 features of the URL and source code. The results confirm that with the use of this system, an accuracy rate of 99.09% was calculated. In [27], a neural network-based classifier method was proposed to detect pages and phishing websites using the Monte Carlo algorithm and the threat reduction principle. In [28] ,the focus was on the impact of training func- tions on neural networks to increase the performance of recommendations.

In [29], four different classes were specified: email header, URL in body, HTML body and main text. The classification is performed in machine learning using 50 features in these categories. The result confirms the accuracy is 98.6%.

In [30], Principal component analysis (PCA) and ran- dom forest (RF) were used to detect zero-day phishing emails. PCA was able to identify zero-day phishing with an accuracy of 99.55%, while RF was able to identify zero day phishing with an accuracy rate of 100%. In [31], the textual content of the email was analyzed and classified. In [32], classification was performed using TF-IDF, man- ual features, and both, as well as 35 features. In this study, the detection rate of phishing attacks was compared with the help of using 6 different algorithms. The Random Forest algorithm had the best end result, with an accuracy rate of 99.55%.

III INVESTIGATION ON RESEARCH GAPS

It has been long understood that phishing is a special- ized social engineering attack in which attackers skillfully use fake emails or websites to trick victims into sharing personal and sensitive facts. There is a need to look at the psychology of people online, whether they are concerned about security when they have the power to change security features. There is a lot of educational material on security and phishing.

There is a large research gap between research and business "on the real positive side". While educational and literary studies mainly specialize in machine learning and heuristics, assuming that positive results are truly excellent, those true positives are sometimes high false positives. Therefore, these discoveries are the best for the ability to recognize phishing sites that have never been encountered before. But, the blacklists fail to generalize to the future unseen instances and also are probably gradual in responding to zero-hour attacks.

IV CONCLUSION AND FUTURE WORK

This survey paper provides three important research elements, an in-depth look at the crime of fraud, a brief review of anti-phishing methods provided by other stud- ies. And a brief survey of research gaps. Scams will never be eliminated, however It is important to understand this before suggesting a solution. Here we have covered the different characteristics of phishing attacks and different strategies for finding phishing sites.

Future work will be to collect studies on the improve- ment of phishing detection systems specifically against phishing websites that are considered the most common means of attack. For greater correct outcomes, Instead of

DOI: 10.17148/IJARCCE.2022.11552

Black-White List approach Naive Bayesian approach, can test and use Artificial Neural Network or Random Forest Classifiers with best hyper-parameters on huge diversified datasets. During testing training or when detecting URLs, Minimum Time to Detection with Maximum Accuracy is a limitation.Future research directions also include the creation of a Chrome browser extension that can detect all the phishing URLs of the current DOM [Document Object Model] page using the recommended machine learning model. This detection engine will also help protect cus- tomers from phishing attacks inside unsafe environments.

ACKNOWLEDGMENT

We would like to thank Prof. D. M. Kanade for his valu-able comments and suggestions to improve the quality of the paper.We would also like to thank the Department of Computer Engineering, K. K. Wagh Institute of Engineer- ing Education and Research, Nashik .

REFERENCES

[1] Anti-phishing Working Group (APWG) Phishing Activity Trends Report 4th quarter 2020, https://docs.apwg.org/reports/apwg trends report q4 2020.pdf
[2] https://atlasvpn.com/blog/a-record-2-million-phishing-sites- reported-in-2020-highest-in-a-decade
[3] Verizon 2020 Data Breach Investigation Report, https://enterprise.verizon.com/resources/reports/2020- databreachinvestigations- report.pdf
[4] World Health Organization, Communicating for Health, Cyber Se- curity, https://www.who.int/about/communications/cyber-security
[5] KasperskyLab (2017) Kaspersky lab:spam and phishing report 2017. https://securelist.com/spam-and-phishing-in- 2017/83833/.Accessed 20 Sept 2018.
[6] CatchPhish: detection of phishing websites by inspecting URLs
[7] A Review on Detecting Phishing URLs using Clustering Algorithms
[8] A. K. Jain, and B. B. Gupta, “Towards detection of phishing websites on client-side using machine learning based approach.” Telecom- munication Systems, vol. 68 no. 4, 2018, pp. 687-700.
[9] Phishing website detection using support vector machines and nature-inspired optimization algorithms
[10] J. Mao, W. Tian, P. Li, T. Wei and Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," in IEEE Access, vol. 5, pp. 17020-17030, 2017, 24-07-2019
[11] J. Mao, W. Tian, P. Li, T. Wei and Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," in IEEE Access, vol. 5, pp. 17020-17030, 2017, 24-07-2019
[12] Phishing website detection using support vector machines and nature-inspired optimization algorithms
[13] Shirazi H, Bezawada B, Ray I (2018) Kn0w thy doma1n name: unbiased phishing detection using domain name based features. In: Proceedings of the 23nd ACM on symposium on access control models and technologies, ACM, pp 69–75
[14] Rao RS, Pais AR (2018) Detection of phishing websites using an of cient feature-based machine learning framework. Neural Compute. Appl. https://doi.org/10.1007/s00521-017-3305-0
[15] Marchal S, Armano G, Gröndahl T, Saari K, Singh N, Asokan N (2017) Of-the-Hook: an efcient and usable client-side phishing prevention application. IEEE Trans Comput 66(10):1717–1733
[16] Jain AK, Gupta BB (2018) A machine learning based approach for phishing detection using hyperlinks information. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-018-0798-z
[17] Y. Zhang, J. I. Hong, and L. F. Cranor, “Cantina,a content based approach to detecting phishing web sites” Proceedings of the 16th international conference on World Wide Web - WWW 07, pp. 639-648
[18] C. L. Tan, K. L. Chiew, K. Wong, and S. N. Sze, “PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder,” Decision Support Systems, vol. 88, pp. 18–27, 2016.

DOI: 10.17148/IJARCCE.2022.11552

[19] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," 2011 Proceedings IEEE INFOCOM, pp. 191-195, 2011.

[20] R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," Journal of Network and Computer Applications, vol. 36, no.1, pp. 324–335, 2013.

[21] S. C. Jeeva and E. B. Rajsingh, "Intelligent phishing url detection using association rule mining," Human-centric Computing and Information Sciences, vol. 6, no. 1, Oct. 2016.

[22] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear re- gression strategy for detecting phishing websites," Soft Computing, vol. 23, no. 12, pp. 4315–4327, 2018.

[23] E. Buber, B. Diri, and O. K. Sahingoz, "Detecting phishing attacks from URL by using NLP techniques," 2017 International Conference on Computer Science and Engineering (UBMK), pp. 337-342, 2017.

[24] E. Buber, B. Diri, and O. K. Sahingoz, "NLP Based Phishing Attack Detection from URLs," Advances in Intelligent Systems and Com- puting Intelligent Systems Design and Applications, pp. 608–618, 2018.

[25] R. M. Mohammad, F. Thabtah, and L. Mccluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, no. 2, pp. 443–458, 2013

[26] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," Telecom- munication Systems, vol. 68, no. 4, pp. 687–700, 2017.

[27] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han J. Wang, "The application of a novel neural network in the detection of phishing websites," Journal of Ambient Intelligence and Humanized Computing, pp 1-15, 2018.

[28] G. Karatas and O. K. Sahingoz, "Neural network based intrusion detection systems with different training functions," 2018 6th In- ternational Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-6, doi: 10.1109/ISDFS.2018.8355327.

[29] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforce- ment learning," Decision Support Systems, vol. 107, pp. 88–102, 2018.

[30] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," Neural Computing and Applications, vol. 31, no. 8, pp. 3851–3873, Jun. 2018.

[31] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 300-301, 2018.

[32] R. S. Rao, T. Vaishnavi, and A. R. Pais, "CatchPhish: detection of phishing websites by inspecting URLs," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 2, pp. 813–825, Oct. 2019.

