International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 8, Issue 2, pp 133-137, February-2024
https://doi.org/10.47001/IRJIET/2024.802019

Detection of Phishing Websites by Using Machine Learning-Based URL Analysis

[1]Shamna Jabbin P, [2]Prof. P. Gopika

[1]PG Student, Dept. of Computer Science and Engineering, EASA College of Engineering and Technology, Tamilnadu, India
[2]Professor, Dept. of Computer Science and Engineering, EASA College of Engineering and Technology, Tamilnadu, India

***Abstract* - In recent years, advancements in Internet and cloud technologies have led to a significant increase in electronic trading in which consumers make online purchases and transactions. This growth leads to unauthorized access to users' sensitive information and damages the resources of an enterprise. Phishing is one of the familiar attacks that trick users to access malicious content and gain their information. In terms of website interface and uniform resource locator (URL), most phishing webpages look identical to the actual webpages. Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested. However, due to inefficient security technologies, there is an exponential increase in the number of victims. The anonymous and uncontrollable framework of the Internet is more vulnerable to phishing attacks. Existing research works show that the performance of the phishing detection system is limited. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the author proposed a URL detection technique based on machine learning approaches. A recurrent neural network method is employed to detect phishing URL. Researcher evaluated the proposed method with 7900 malicious and 5800 legitimate sites, respectively. The experiments' outcome shows that the proposed method's performance is better than the recent approaches in malicious URL detection. In recent years, with the increasing use of mobile devices, there is a growing trend to move almost all real-world operations to the cyber world. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. The experimental results depict that the proposed models have an outstanding performance with a success rate.**

Keywords: Phishing, Phishing Attack, Machine Learning, Network Attack.

I. INTRODUCTION

Phishing is a fraudulent technique that uses social and technological tricks to steal customer identification and financial credentials. Social media systems use spoofed e-mails from legitimate companies and agencies to enable users to use fake websites to divulge financial details like usernames and passwords. Hackers install malicious software on computers to steal credentials, often using systems to intercept username and passwords of consumers' online accounts. Phishers use multiple methods, including email, Uniform Resource Locators (URL), instant messages, forum postings, telephone calls, and text messages to steal user information. The structure of phishing content is similar to the original content and trick users to access the content in order to obtain their sensitive data. The primary objective of phishing is to gain certain personal information for financial gain or use of identity theft. Phishing attacks are causing severe economic damage around the world. Moreover, most phishing attacks target financial/payment institutions and webmail, according to the Anti-Phishing Working Group (APWG) latest Phishing pattern studies.

In order to receive confidential data, criminals develop unauthorized replicas of a real website and email, typically from a financial institution or other organization dealing with financial data. This e-mail is rendered using a legitimate company's logos and slogans. The design and structure of HTML allow copying of images or an entire website. Also, it is one of the factors for the rapid growth of Internet as a communication medium, and enables the misuse of brands, trademarks and other company identifiers that customers rely on as authentication mechanisms. To trap users, Phisher sends "spooled" mails to as many people as possible. When these e-mails are opened, the customers tend to be diverted from the legitimate entity to a spoofed website.

Phishing is the most commonly used social engineering and cyber attack. Through such attacks, the phisher targets naïve online users by tricking the mint revealing confidential information, with the purpose of using it fraudulently. In order to avoid getting phished, users should have awareness of phishing websites. Have a blacklist of phishing websites which requires the knowledge of website being detected as phishing. Detect them in their early appearance, using machine learning and deep neural network algorithms of the above three, the machine learning based method is proven to be most effective than the other methods. Even then, online

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 2, pp 133-137, February-2024

https://doi.org/10.47001/IRJIET/2024.802019

users are still being trapped into revealing sensitive information in phishing websites. A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and web pages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared. The phishing website has evolved as a major cyber security threat in recent times. The phishing websites host spam, malware, ransom ware, drive-by exploits, etc. A phishing website many a time look-alike a very popular website and lure an unsuspecting user to fall victim to the trap. The victim of the scams incurs a monetary loss, loss of private information and loss of reputation. Hence, it is imperative to find a solution that could mitigate such security threats inatimely manner. Traditionally, the detection of phishing websites is done using blacklists. There are many popular websites which host a list of blacklisted websites, e. g. Phis Tank. The blacklisting technique lack in two aspects, blacklists might not be exhaustive and do not detect a newly generated phishing website.

II. PROBLEM FORMULATION

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the project is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each. algorithm. Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion .Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. Only data mining is not sufficient the problem of data overloaded.

III. BASIC IDEA OF OUR SCHEME

In recent times machine learning techniques have been used in the classification and detection of phishing websites. In, this paper we have compared different machine learning techniques for the phishing website. In our daily life, we carry out most of our work on digital platforms. Using a computer and the internet in many areas facilitates our business and private life. It allows us to complete our transaction and operations quickly in areas such as trade, health, education, communication, banking, aviation, research, engineering, entertainment, and public services. The users who need to access a local network have been able to easily connect to the Internet anywhere and anytime with the development of mobile and wireless technologies. Although this situation provides great convenience, it has revealed serious deficits in terms of information security. Thus, the need for users in cyberspace to take measures against possible cyber-attacks has emerged. Intrusion detection is the demonstration of recognizing undesirable movement on a system or a gadget.

The method of reaching target users in phishing attacks has continuously increased since the last decade. This method has been carried out in the 1990s as an algorithm-based, in the early2000s based on e-mail, then as Domain Spoofing and in recent years via HTTPs. Due to the size of the mass attacked in recent years, the cost and effect of the attacks on the users have been high. The average financial cost of the data breach as part of the phishing attacks in 2019 is $ 3.86million, and the approximate cost of the BEC (Business Email Compromise) phrases is estimated to be around $ 12 billion. Also, it is known that about 15% of people who are attacked are at least one more target. With this result, it can be said that phishing attacks will continue to being carried out in the ongoing years. Figure 1 also supports this idea and show the number of phishing sites in2019, and as can be seen from it, there is an increasing trend in this type of attack. In this regard, regular reports published by APWG (Anti Phishing Working Group) are an important guide for the researchers. According to the reports, the number of phishing sites is reached to approximately 640,000 sites were determined in 2018, and in the first three quarters of 2019, this number was reported as 629,611.Reportsfor the last quarter of 2019havenotbeen published yet. However, it can be said that the phishing attacks not only continue, but also there will be an increase in the number of attack types compared to the previous year.

This increase indicates that phishing attacks are used more by attackers. Because they are easy to design. Phishing

ISSN (online): 2581-3048
Volume 8, Issue 2, pp 133-137, February-2024
https://doi.org/10.47001/IRJIET/2024.802019

attacks are based on the attacker's creation of a fake website, as depicted in Figure 2. First, a phisher makes fake websites, including a phishing kit. Then, the victim is directed to the fake website with the prepared email. Believing that the e-mail and URL are secure, the victim uses the fake website by clicking on the URL.

After this moment, the Phishing kit receives the victim's credentials and sends it to the phisher. Finally, Phisher makes fake earning from the legitimate website using the victim's credentials. These sites generally have very similar or even identical visuals. In an e-mail that is thought to be sent from a trusted source, the target is directed to this fake web site. The target In this way, the attacker gets information and / or earnings. Reliable e-mail contents are created in different ways for the victim to believe. Previously, e-mails with low probability offers, urgent texts, links, or attachments that may be relevant and unusual senders were used. Today, reliable organizations or similar links to these organizations are preferred. Attackers prefer reaching to victims by using a secure communication protocol, and the real URL is served by changing in a way that is close to the original. At this stage, if the victim knows the website is fake, he can protect himself from the attack. It is very difficult for the victim to detect the attack by himself, because mainly this type of messages gave some alert messages to the users, and aims to make panic for entering his confidential data to the forwarded page.

IV. RELATED WORK

1) Ebubekir Buber, Onder Demi, Ozgur Koray Sahingoz, "Feature selections for the machine learning based detection of phishing websites".

Phishing websites are malicious sites which impersonate as legitimate web pages and they aim to reveal users important information such as user id, password, and credit card information. Detection of these phishing sites is a very challenging problem because phishing is mainly a semantics based attack, which especially abuses human vulnerabilities, however not network or system vulnerabilities.

2) Jimmy Moedjahedy, Arief Setyanto, Fawaz Khaled Alarfaj, Mohammed Alreshoodi", Combine Correlation Features Selection for Detecting Phishing Websites Using Machine Learning".

Internet users are continually exposed to phishing as cybercrime in the 21st century. The objective of phishing is to obtain sensitive information by deceiving a target and using the information for financial gain. The information may include a login detail, password, date of birth, credit card number, bank account number, and family-related information.

3) Priya Saravanan a, Selvakumar Subramanian b, "A Framework for Detecting Phishing Websites using GA based Feature Selection and ARTMAP based Website Classification".

Nowadays, Phishing attack has gained more attention among all the other attacks existing in online social media. The fraudulent E-mail sent form the fake website that looks like the legitimate website is the initial carter for launching the phishing attacks. This is a kind of social engineering attack in which, the user is targeted for stealing the personal information, viz., user name, password, and banking credentials for committing the financial crimes.

4) Altyeb Taha," Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting".

The continuous development of network technologies plays a major role in increasing the utilization of these technologies in many aspects of our lives, including e-commerce, electronic banking, social media, e-health, an de-learning. In recent times, phishing websites have emerged as a major cyber security threat. Phishing websites are fake web pages that are created by hackers to mimic the web pages of real websites to deceive people and steal their private information, such as account usernames and passwords.

5) Ye Cao, WeiliHan, "Anti-phishing based on Automated Individual White-List".

In phishing and pharming, users could be easily tricked into submitting their username/passwords into fraudulent web sites whose appearances look similar as the genuine ones. The traditional blacklist approach for anti-phishing is partially effective due to its partial list of global phishing sites. In this paper, we present a novel anti phishing approach named Automated Individual White-List

V. CONCLUSION

In this project, we have explored how well to classify phishing URLs from the given set of URLs containing benign and phishing URLs. We have also discussed the randomization of the dataset, feature engineering, feature extraction using lexical analysis host-based features and statistical analysis. We have also used different classifiers for the comparative study and found that the findings are almost consistent across the different classifiers. We also observed dataset randomization yielded a great optimization and the accuracy of the classifier improved significantly. We have adopted a simple approach to extract the features from the URLs using simple regular expressions. There could be more features that can be experimented and that might lead to improving further the accuracy of the system. The dataset used in this paper contains the URLs list which may be a little old,

ISSN (online): 2581-3048
Volume 8, Issue 2, pp 133-137, February-2024
https://doi.org/10.47001/IRJIET/2024.802019

hence regular continuous training along with a new dataset would enhance the model accuracy and performance significantly. In our experiment we have not used the content based features as the main problem with the content-based strategy for detecting phishing URLs is the non-availability of phishing web-sites and the life span of the phishing website is small, and it is difficult to train an ML classifier based on its content-based features. In the future, we would like to incorporate a rule-based prediction based on the content analysis of a URL. Hence, the combination of classification based lexical analyzer along with a rule-based URL content analyzer for phishing URL detection would provide a comprehensive solution.

VI. FUTURE WORK

In recent years, due to the evolving technologies on networking not only for traditional web applications but also for mobile and social networking tools, phishing attacks have become one of the important threats in cyberspace. Although most of security attacks target on system vulnerabilities, phishing exploits the vulnerabilities of the human end-users. Therefore, the main defense form for the companies is informing the employees about this type of attack. However, security managers can get some additional protection mechanism which can be executed either decision support system for the user or as a prevention mechanism on the servers. In this paper, we aimed to implement a phishing detection system by using some more machine learning algorithms and work workout in AI.

REFERENCES

[1] State of Cybersecurity Implications for 2016. An ISACA and RSA Conference Survey.[Online]. Available https://cybersecurity.isaca.org/csx-resources/state-of- ybersecurityimplications-for-2016.[Accessed: 09-Mar-2020].

[2] Republic of Turkey, "National Cyber Security Strategy, 2016," Ministry of Transport Martime Affairs and Communications.

[3] R. Loftus, "What cybersecurity trends should you look out for in 2020?," Daily English Global blogkasperskycom. [Online]. Available: https://www.kaspersky.com/blog/secure-futures-magazine/2020- cyber security-predictions/32068/. [Accessed: 09-Mar-2020].

[4] E. Buber, Ö. Demir and O. K. Sahingoz, "Feature selections for the machine learning based detection of phishing websites," 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, 2017, pp. 1-5.

[5] "Retruster," Retruster. [Online]. Available: https://retruster.com/blog/2019-phishing-and-email-fraud-statistics.html. [Accessed: 09-Mar-2020].

[6] "Phishing Activity Trends Reports, 1st-2nd-3rd Half" APWG. [Online]. Available: https://apwg.org/trendsreports/. [Accessed: 09-Mar-2020].

[7] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," Proceedings of the 4th ACM workshop on Digital identity management - DIM 08, pp. 51–60, 2008.

[8] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840–843, 2008.

[9] M. Khonji, Y. Iraqi, and A. Jones, "Phishing Detection: A Literature Survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

[10] 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, 2007.

[11] S Sugumaran, B Buvaneswari, KS Senthil Kumar "Optimization Based collision Avoidance in under water wireless sensor Network2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), pages 1-6, Publishers IEEE.

[12] D Nethra Pingala Suthishni, GP Ramesh Kumar intrusion detection analysis by implementing fuzzy logic, journal Int. J. Appl. Eng. Res volume 11, issue 5, page 3216-3220.

[13] Dr. E. CHANDRA BLESSIE, S.GNANAPRIYA "A REVIEW ON OPTIMIZATION FOR PRE-PROCESSING TECHNIQUES IN DATA MINING", The International journal of analytical and experimental modal analysis, ISSN NO: 0886-9367.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 2, pp 133-137, February-2024

https://doi.org/10.47001/IRJIET/2024.802019

Citation of this Article:

Shamna Jabbin P, Prof. P. Gopika, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 2, pp 133-137, February 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.802019

****

