International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST52310210

A Novel Approach on Detecting Phishing Attacks on URLS using ML Techniques

M. Nirmala[1], I. Lakshmi Prasanna[2], B. Meghana Sai Rajeshwari[3], Ch. Gopi[4], A. Vidyadhari[5]

[1]Assistant Professor, Department of Information Technology, Kallam Haranadha Reddy Institute of Technology, Chowdavaram, Guntur (Dt), Andhra Pradesh, India

[2]B. Tech Students, Department of Information Technology, Kallam Haranadha Reddy Institute of Technology, Chowdavaram, Guntur (Dt), Andhra Pradesh, India

ARTICLEINFO

Article History:
Accepted: 01 March 2023
Published: 13 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
207-212

ABSTRACT

The simplest way to obtain sensitive information from unwitting users is through a phishing attack. The phishers' goal is to obtain sensitive information such as usernames, passwords, and bank account information. The proposed system is primarily concerned with detecting and preventing phishing websites. The websites are discovered using a web crawler. Phishing sites that are detected are added to the blacklist. The blacklist only contains fake websites. Web crawling is concerned with obtaining the web page's links. A phishing website can usually be identified by its URL and HTML code. The check website page alerts users to phishing websites and helps them avoid becoming victims of such attacks. This software is extremely useful in identifying and preventing the PHISHING.

Keywords: HTML, PHISHING, HTML

I. INTRODUCTION

Phishing is a new term derived from the word "fishing," and it refers to the act by which attackers entice users to visit a faked Web site. It is a new type of network attack in which the attacker copies content from a well-known company or bank's website and creates a phishing website. To attract more users, the attacker keeps the phishing website visually similar to the corresponding legitimate website. The phishers must duplicate the target site's content and use tools to (automatically) download the target site's Web pages. The proposed system assists the user in detecting and determine the phishing websites. It employs a web crawler to crawl the website's hyperlinks. The website is correct. It employs a web crawler to crawl the website's hyperlinks. If there are no hyperlinks on the website, it will be added to the blacklist. Those websites are forgeries, and the user does not want to visit them. The web crawler can validate the user's URL; if the result is phished, it warns the user and adds the website to a blacklist. The user can scan the website;

if it is a phishing site, an alert message will be displayed to warn the user.

II. LITERATURE SURVEY

A. Antiphishing to Keep Users Safe from Phishing

AntiPhish is used to prevent users from visiting fraudulent websites, which may result in a phishing attack. AntiPhish tracks the sensitive information that the user enters and alerts the user whenever he or she attempts to share his or her information on an untrusted website. The most effective explanation is to encourage users to only visit trusted websites. This approach, however, is implausible. In any case, the user may be duped. To overcome the problem of phishing, it is therefore mandatory for the associates to present such explanations. Creepy websites are widely accepted alternatives for identifying "clones" and maintaining phishing website records in the hit list.

B. Learning to Spot Phishing Emails

An alternative to detecting these attacks is the required process of system reliability on a trait intended to reflect the besieged deception of users via electronic communication. This method can be used to detect phishing websites or text messages sent via email to lure victims. To date, approximately 800 phishing emails and 7,000 non-phishing emails have been traced, with over 95 percent of them correctly identified based on 0.09 percent of real emails along with categorization. We should simply conclude with deception detection methods and the changing nature of attacks.

C. E-banking phishing detection method using fuzzy data mining

Identifying and classifying phishing websites, which are primarily used for e-banking services, is a difficult and dynamic task. Because of the presence of various ambiguities in the identification, some critical data mining techniques can prove an effective way to keep e-commerce websites secure, as it deals with the consideration of different quality variables rather than precise values. In this paper, an intelligent, resilient, and successful model is used to detect e-banking phishing websites by resolving "fuzziness" in the evaluation of the e-banking phishing website. The implemented model is based on fuzzy logics and data mining algorithms to consider various successful factors of the e-banking phishing website.

D. Dynamic Feature Extraction from Entered URL WHOIS Database:

The life of phishing site is very short, therefore; this DNS information may not be available after some time. If the DNS record is not available anywhere then the website is phishing. If the domain name of the suspicious webpage is not match with the WHOIS database record, then webpage considered as phishing.

Features considered are listed below:

1. Having an internet protocol (IP) Address If the IP address is used as the domain of the URL, such as http://125.98.2.142/contoh.html, it can be suspected that attempts to steal information.

2. URL Length Long URLs can also be suspected of being a phishing site. If the URL length is greater than or equal to 75 characters then the URL included as a phishing site.

3. Shortening Service: URL shortening is a method in which a URL is made to be shorter, which this domain ill connect to the web page that has a URL that is longer such as, http://sekolah.ini.ac.id/ URLs can be shortened to "bit.ly/21FXWl5".

4. Having @ Symbol: URLs using the symbol @ will lead to the browser to ignore everything that precedes the @ symbol.

5. Double Slash Redirecting: Double slash or "//" indicates that the user will be redirected to another site. The position of the use of double slash usually appears at the sixth position as written at this link http://amikom.ac.id. 6. Prefix or Suffix: Rarely a legitimate URL using symbols dashboard, but phisher will add a prefix or suffix to be separated by (-) in the domain name, so the user will think to have a

legitimate access to sites such as http://www.amikom-keren.com.

7. Having Sub Domains The domain name may have a code for each country (cc TLD) such as "id", or for an academic educational institution "ac" and combined "ac.id" or also called two-level domain (SLD). Stages for extracting the feature of the first to do is remove the "www" in the URL and remove cc DTL if any. Then calculate the remaining dots, if the number of points is greater than one, then the URL can be classified as "suspect" because of only the subdomain. However, if the number of points greater than the two it will be categorized as a phishing because it has several subdomains, and sites categorized as legitimate if it does not have a subdomain.

8. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer): HTTPS is an essential component in a site that looks at legality.

9. Domain Registration Length: Categorized as phishing sites are valid only in a short time and are used for a single year.

10. Favicon: Favicon is an image used as an icon on a website, favicon also indicates the identity of the website. However, if the favicon is displayed apart in the address bar, it can be suspected that the website is a phishing website.

11. Port: Port used to validate certain services such as HTTP. The use of a firewall, proxy, and Network Address Translation or NAT can perform automatic blocking and can be opened in accordance with the wishes. But if all the ports are opened, then the phisher will find loopholes and enable any desired services such as stealing information.

12. HTTPS Token: In general, the https token can be added by phisher on the domain URL and has the objective to distract the user like at http://httpswww-amikom-coolest-college.com/

13. Request URL On the website is legal, website addresses, pictures, videos, and sounds contained on the web page with the same domain and does not take away from another domain.

14. Abnormal URL: On the website is legitimate, then the identity of the website will be contained in the URL. 15. Links in Tags: Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use tags to offer metadata about the HTML document.

16. SFH: SFHs that contains an empty string or “about:blank” are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.

17. Submitting to Email: The official website will generally send personal information to the server for processing. While the phisher will be sending the information to his personal email, it can be suspected by the use of scripts on the server side functions such as "mail ()" and on the client side will use the mailto().

18. Abnormal URL: This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL.

19. Redirect: The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times.

20. On_mouseover: Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, we must dig-out the webpage source code, particularly the “onMouseOver” event, and check if it makes any changes on the status bar.

21. Right Click: Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. This feature is treated exactly as “Using onMouseOver to hide the Link”. Nonetheless, for this feature, we will search for event

"event.button==2" in the webpage source code and check if the right click is disabled.

22. Popup Window: It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

23. Iframe: IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the "iframe" tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the "frameBorder" attribute which causes the browser to render a visual delineation.

24. Age of Domain: This feature can be extracted from WHOIS database. Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6 months. 25. DNS Record: For phishing websites, either the claimed identity is not recognized by the WHOIS database or no records founded for the hostname. If the DNS record is empty or not found then the website is classified as "Phishing", otherwise it is classified as "Legitimate".

26. Web traffic: This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period of time, they may not be recognized by the Alexa database. By reviewing our dataset, we find that in worst scenarios, legitimate websites ranked among the top 100,000. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as "Phishing". Otherwise, it is classified as "Suspicious".

27. Page Rank: PageRank is a value ranging from "0" to "1". PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage. In our datasets, we find that about 95% of phishing webpages have no PageRank. Moreover, we find that the remaining 5% of phishing webpages may reach a PageRank value up to "0.2".

28. Google Index: This feature examines whether a website is in Google's index or not. When a site is indexed by Google, it is displayed on search results (Webmaster resources, 2014). Usually, phishing webpages are merely accessible for a short period and as a result, many phishing webpages may not be found on the Google index.

29. Links Pointing to Page: The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain. In our datasets and due to its short life span, we find that 98% of phishing dataset items have no links pointing to them. On the other hand, legitimate websites have at least 2 external links pointing to them.

III. EXISTING SYSTEM

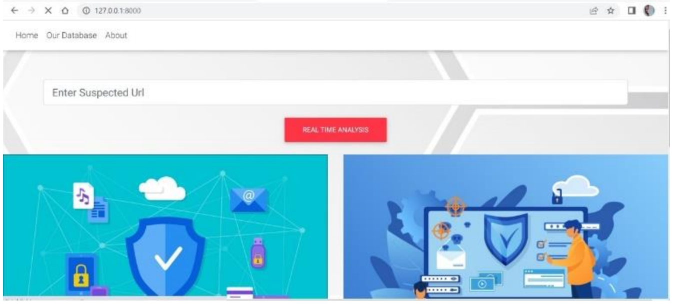
The existing system manually determines whether a site is a phishing site or not, but finding those phishing sites is difficult. A legal Web site's Webmaster scans the root DNS for suspicious sites on a regular basis (e.g.www.1cbc.com.cnvs.www.icbc.com.cn). Because the phisher must duplicate the target site's content, he must use tools to (automatically) download the target site's Web pages. This type of download can thus be detected at the Web server and traced back to the phisher. DNS scanning increases the overhead of DNS systems and may cause problems with normal DNS queries; additionally, many phishing attacks do not require a DNS name. Smart phishers can easily create tools.

IV. PROPOSED SYSTEM

The proposed system uses a web crawler to detect phishing websites. The search begins by crawling your site's pages. Then it proceeds to visit the links (web page addresses or URLs) on your site. In our proposed system, the administrator can login and enter any website URL into the web crawler, which then searches the URL and identifies the page's hyperlinks. The links are visible on the current page. If there are no hyperlinks on the website, it will be added to the blacklist. Phishing sites are included on the blacklist. Those websites are forgeries, and the user does not want to visit them. The user can now enter the URL of a website into the check website page, and then checks whether the site is phished or not.

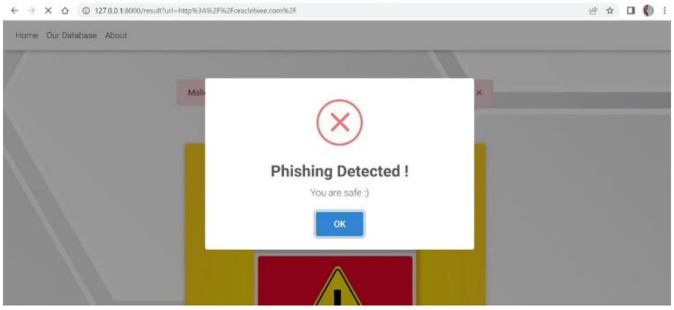
V. SYSTEM IMPLEMENTATION

Implementation is the process of turning a theoretical design into a working system. As a result, it may be the most crucial stage in achieving a successful new system and in instilling confidence in the user that the new system will work and be effective. The implementation stage entails careful planning, investigation of the existing system and its implementation constraints, design of changeover methods, and evaluation of changeover methods. The process of putting a new system design into operation is known as implementation. It is the phase of the installation process that focuses on user training, site preparation, and file conversion. The most important factor to consider here is that the conversion should not disrupt the organization's operations.



VI. RESULTS

Authorized users can access data by logging on this website as shown in Fig. 1. On entering the registered user ID and password, it goes to the webpage where the user should enter the URL of the webpage and enter the scan button. Once this is done the web crawler will scan the URL and notify the user whether the sight is phishing or not. So, this makes the user's details secured from the phishing website.


Phishing Detected !

You are safe :)

VII. CONCLUSION AND FUTURE WORK

There is no single solution to phishing. It is a critical situation in which phishers are constantly attempting to devise novel methods of manipulating consumers. Online users should engage in regular risk assessment to detect new techniques that could lead to a successful Phishing attack. To find safer alternatives, users must be aware of the dangers of advanced malware that exist today. Furthermore, security teams must implement advanced methodologies that can eliminate advanced threats that have recently been bypassed by their predictable resentment. Contributions are also made in detecting identity theft and phishing emails. It does not involve in the rising trends towards e-mail outsourcing. Log analysis and communication taking place across.

VIII. ACKNOWLEDGEMENT

We express our heartfelt gratitude to our Guide Mrs. Nirmala, Department of Information Technology, for her guidance, inspiration, and suggestions in our quest for knowledge. We would like to thank our parents for their tremendous contribution in assisting us in reaching this point in our lives. This would not have been possible without their constant and selfless love, cooperation, and encouragement. However, it would not have been possible without the generous support and assistance of many people. We would like to express our heartfelt gratitude to each and every one of them.

IX. REFERENCES

[1]. Routhu Srinivasa Rao and Syed Taqi Ali, 2015 “PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach” Department of Computer Engineering, National Institute of Technology, Kurukshetra , Haryana, India, Procedia Computer Science 54,(147 – 156).

[2]. Hemali Sampat, Manisha Saharkar, Ajay Pandey, Hezal Lopes, 2018” Detection of Phishing Website Using Machine Learning” Department of Computer Engineering, Universal College of Engineering, Vasai , Maharashtra, India,IRJET Volume-5 Issue-3.

[3]. S. Carolin Jeeva and Elijah Blessing Raj singh. 2016 “Intelligent phishing url detection using association rule mining” Department of Computer Applications ,Karunya University, Coimbatore, India.

[4]. Sa'id Abdullah Al-Saaidah, 2017 “Detecting Phishing Emails Using Machine Learning Techniques” Department of Computer Science Faculty of Information Technology Middle East University.

[5]. Ram B. Basnet1, Andrew H. Sung, Quingzhong Liu, 2014 “Learning to Detect Phishing URLs” Colorado Mesa University, 1100 North Ave. Grand Jct. CO 81501, USA IJRET.

[6]. Ankith Kumar jain and BB Guptha,2017 “Phishing Detection: Analysis of Visual Similarity Based Approaches” National Institute of Technology, Kurukshetra, India, Hindawi Security and Communication NetworksVolume 2017.

[7]. Jian Mao, Jingdong Bian, Wenqian Tian, Shishi Zhu, Tao Wei, Aili Li and Zhenkai Liang, 2018 “Detecting Phishing Websites via Aggregation Analysis of Page Layouts” Procedia Computer Science 129 (2018) 224–230. [8] R. Kiruthiga, D. Akila,2019 “Phishing Websites Detection Using Machine Learning” (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11.

[8]. Moitrayee Chatterjee and Akbar Siami Namin, 2019“Detecting Phishing Websites through Deep Reinforcement Learning”IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). [10] M E Pratiwi ,2018 “Phishing Site Detection Analysis Using Artificial Neural Network” Journal of Physics: Conference Series1140 (2018) doi:10.1088/1742-6596/1140/1/012048.

Cite this article as :

M. Nirmala, I. Lakshmi Prasanna, B. Meghana Sai Rajeshwari, Ch. Gopi, A. Vidyadhari, " A Novel Approach on Detecting Phishing Attacks on URLS using ML Techniques", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X, Volume 10, Issue 2, pp.207-212, March-April-2023. Available at doi : https://doi.org/10.32628/IJSRST52310210
Journal URL : https://ijsrst.com/IJSRST52310210