URL Based Phishing Website Detector

[1]Kishor Shanmugavel, [2]Soumik Rakshit, [3]Barathvaraj M , [4]Sangeetha. D, [5]Umamaheswari. S

[1,2,3]Students ,Department of Information Technology, Madras Institute of Technology, Anna University, Chennai

[4]Assistant Professor ,Department of Information Technology, Madras Institute of Technology, Anna University, Chennai

[5]Associate Professor ,Department of Information Technology, Madras Institute of Technology, Anna University, Chennai

Corresponding Author Email:-***projectdarkinblade@gmail.com***

DOI: 10.29322/IJSRP.12.05.2022.p12529
http://dx.doi.org/10.29322/IJSRP.12.05.2022.p12529

Paper Received Date: 21st April 2022
Paper Acceptance Date: 6th May 2022
Paper Publication Date: 14th May 2022

Abstract:-

In the current scenario most popular cyber crime today is phishing, Phishing is a social engineering-based attack where the phisher targets to retrieve the user's legitimate confidential details to exploit and it is commonly done through sending mail or any other electronic communication in an automated fashion. In this paper a Machine Learning based approach on URLs of a website is done to detect phishing websites aiming to reduce the number of features for the model and improve the accuracy. The data-set contains numerous tested URLs which are labeled as good and bad. An URL is analyzed completely broken into tokens and with classified datasets, these tokens are pooled in two categories accordingly and aggregated. This model is strapped with a web application and web extension which gets input as URL and do a prediction test to give the result.The previous models which uses machine learning model doesn't utilize the patterns of URL in the structure instead other parametric data which produce high number of features increasing the computation process. The accuracy of those models come to 93%. Sometimes those data fail to be consistent. By targeting the structure of the URL to identify the pattern, the number of feature gets reduced which will wont have the need of high computation. The prescribed model in this paper produces uses logistic regression which produced an accuracy up to 96% and contains fewer features compared to previous models. This model is strapped with a web application and web extension which acts as front-end interface and it has successfully defended an email-based phishing attack simulation in detecting a phishing website.

Keywords- Phishing, Phishing Detection,Detection Tool, Social Engineering, URLSynthesis, Machine Learning

1.INTRODUCTION

Nowadays with the ease of accessto the Internet, people carry out so many activities like online shopping, online billpayment, etc. There are many cyber crimes, one among them is phishing. Phishing is a global threat that utilizes social engineering with website spoofing. This threat not only causes tremendous financial losses to the users but also harms reputation and trust. Consequently, this affects the legitimate websites which are mimicked to carry out phishing attacks. Anattacker uses social skills to obtain or compromise information about a target that an attacker tries to exploit, this is called social engineering.

A typical phishing attack is of threesteps (i)Lure- Gaining user's attention to the manipulative message, (ii)Hook- Deceiving the user to make them respond to the request made by the attacker, (iii)Catch- After the user responds to the message extract all confidential details of the user. One of the major identifications of a phishing website that is modeled to deceive the user is the URL (Uniform Resource Locator). It is a global address for the documents and acts as an identification on the World Wide Web.

An attacker tries to model the URL in sucha way that it is similar to URLs that a user comes across on a day-to-day basis, without raising any alarm the user uses thepage whose main goal is to extract information from the user.

There are many methods to identifya phishing page with a URL and one such method is Machine Learning, which uses datasets of already existing phishing pagesand extracts features to analyze patterns to identify if a particular URL is a phishing page or not. In this paper, we discuss a Machine Learning Model using Natural Language Processing on URLs and build a Web Application and Extension which integrates with the model so it can be usedas a tool. The tool developed will also be tested and verified in an attack simulation to prove its testing.

Target-dependent and Target- independent are two main classifications done to extract key features which can be compared between a legitimate website and a potential phishing website[1]. The target-dependent method measures how different a phishing web page is from a suspected phishing web page. Offline information or information given by the search engine of the web page is collected. The Target-independent method uses generic features from the URL, visual appearance, servers, etc. By using these features a web page can be classified as a phishing page or not. URL contains lexical features like domain, directory, file, hostname, path, arguments, protocol, etc[3]. These are extracted and utilized in the Machine Learning model to give a prediction.Some of the popular techniques are

(i) Blacklist- The URL is stored in a database that contains only phishing websites, if the current URL is found in the database it is flagged. The database needs to be updated from time to time, as new URLs are developed every time to overcome the current methodology of detection. (ii)Heuristic Based- It is an extension of blacklisting, where it examines every single URL which can spot specific characteristics to determine aweb page phishing. (iii) Visual Similarity- It extracts images from legitimate websitesto extract features to do a comparison to identify the phishing website. (iv)MachineLearning- It is a popular method as new variations are developed to address an issue every single time and also to improvethe accuracy[3]. Supervised Learning opts were to train the classifier features of both legitimate and phishing websites which areextracted and trained to give prediction[5]. Some of the classification techniques are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine[2][5].To address the reduction of the number of features where numerous features are utilized to produce a machine learning model so far, the structure of the URL will be synthesized and converted to tokens. These tokens will be categorized and analysed to find the pattern between legitimate websites and malicious websites. The model prescribed in this paper will try to reduce the number of features for the machine learning model and match the accuracy with the existing model which takes in numerous features. The model will also be supported with an interface like web application and web extension and will be tested in detecting phishing attacks, primarily email-based.

2.PROPOSED APPROACH

A machine learning model is developed around the features of URL. To make the computation complexity less, the approach revolves around a small number of features. The best solution is Natural Language Processing (NLP), where it is used for language processing, decomposing the URL to tokens, and analyzing it to obtain a pattern. The data setcontains both malicious and secure websites URL tagged with them, they are decomposed to tokens and grouped to find specific and repeating patterns which will be helpful to detect if a website is malicious or not. It is tested with different classifiers and the one which

gives the highest accuracy is used.

The machine learning model whichis built needs to be utilized as a tool therefore an interface is developed i.e., web application and web extension.

3.ALGORITHM FOR MACHINE LEARNING MODEL:

1. Start
2. Load the data set
3. Utilize regular expressionto kenizer
4. Synthesise the data set and splitit into tokens using snowball stemmer
5. Categories them according tolabel: good or bad
6. Split data to train, test

7. Use logistic regression as aclassifier to train
8. Use a pipeline to automate allprevious steps in an instant
9. Test the accuracy
10. Shift this model in pickle file toexport
11. Enter URL input in a stringvariable
12. Use the variable to predict good orbad

The figure 3.1. shows the web application acts as a front runner where it fetches the prediction from the machine learning model by giving an input to trigger predict function. The input URL is entered manually and the value is submitted as a form, where the string valueis carried to the machine learning model for prediction. The model predicts using the URL and labels it as'[GOOD]' for secure websites and '[BAD]' for maliciouswebsites.

Figure 3.1 . Shows the Web Application

The figure 3.2 . shows the web extension will be mounted in a web browser directly. It'll read the URL directly as input from the tab where users enter different websites. It uses built- in chrome developer function chrome.tabs.get Selected. The input URL is sent to the intermediary server with the help of AJAX which connects the browser extension to the machine learning model. The PHP server sends the input URL by passing it as an argument to the EXEC command which is used for calling the machine learning model. The machine learning model returns the predicted result which is sent back via the server and AJAX is used to render it without the needof reloading the web page. A phishing attack is designedand simulated in real life to get anunderstanding of a real-life attack and analyze the vulnerabilities present to take advantage of.

Figure 3.2.Shows the Web Extension

This is also done to understand the art of mimicking and deceiving people in the world of the internetthrough a simple influencing message to make people respond carelessly. Further, the tool which is developed to detect phishing websites like these used in a phishing attack is tested to see how preciseit can detect.

Algorithm for a PhishingAttack Simulation:

1. Identify fields of interest to mimic
2. Survey interests of people to extracthighlighting characteristics to targetusers using that and attain information about them
3. Gather all data collected and organize
4. Identify a suitable website to mimic
5. Build a website seeming close to theoriginal one

6. Draft the range of users who can beassociated with the website
7. Structure message to send whichlooks original and authentic
8. Prepare details to launch a phishingattack
9. Launch the attack and wait for aresponse
10. Collect all responses received andorganize them
11. Use the response data to exploitfurther

Figure 3 .3 Shows the Launch of a phishing attack.

Figure 3.3 the shows the Launch of a phishing attack, The area of interest which is opted to launch a phishing attack ise-learning, mainly targeting students who use these platforms to learn technologies, subjects, etc from any remote places at anytime.

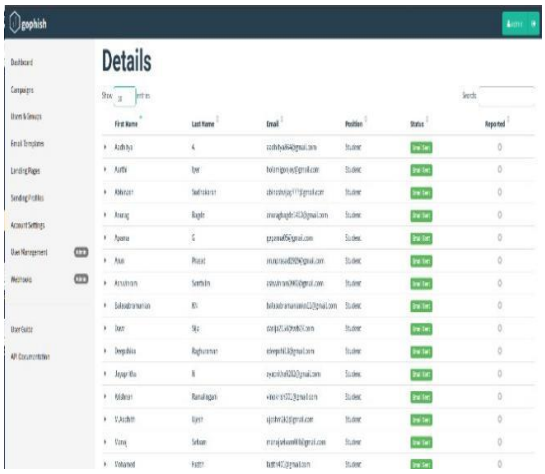
3. RESULTS & DISCUSSION

The accuracy obtained from the machine learning model which uses language processing of the URLs to detectmalicious websites is 96%. The machine learning model is integrated with the web application and web extension. The tools are tested extensively and deployed for use. The web application is hosted in Heroku which is cloud-based. The web application asks the user to enter the URLwhich they want to test to see if the website belonging to the URL is secure ornot. The prediction value from the machine learning model is displayed.The web extension is mounted on the Google Chrome web browser. Whenever the user opens a web page, the extension reads the URL directly and sends it to the machine learning model to detect and givea response. The response is displayed by the web extension.

A phishing attack is launched for simulating a real-life phishing attack. The website chose to mimic as an e-learning website is simplilearn.com. Details have been gathered and organized and used to launch an attack on targets. The medium used is electronic mail with help of go phish, an open-source phishing framework. The response which is received after launching the attack successively is stored in a database which can be used to exploit further. This website is tested with the detection tool and it detects it as a malicious website.

4. CONCLUSION

A high accuracy machine learning model of 96% is built with a smaller number of features to work on. Completelytested to be reliable and solid. The machine learning model is further integrated with a web application and extension to use those functionalities as a tool. A phishing attack simulation is framed and launched successively and captured the data from targets. This model which is used in the phishing attack



simulation is also tested with our tool and it detects precisely.

FUTURE SCOPE

The proposed system is implemented as a detection tool. This can further be improved and support blacklisting URLs that are malicious and record every input given to detect to use them and update the machine learning model with current datasets. This will keepup with evolved URLs which are modeled to escape detection and stay up to date with current attacks. The tool can also be polished to be utilized as a phishing framework.

REFERENCES.

1.Xiuwen Liu and Jianming Fu (2020); SPWalk: Similar Property OrientedFeature Learning for Phishing Detection , IEEE Access.doi:10.1109/ACCESS.2020.2992381.

2.Preeti, Rainu Nandal, and Kamaldeep Joshi,(2021);Phishing URL Detection Using Machine Learning,Advances in Communication and Computational Technology,Springer , vol. 668,pp 547–560,

3.Harshal Tupsamudre, Ajeet Kumar Singh and Sachin Lodha(2019). ‘Everything Is in the Name – A URL Based

Approach for Phishing Detection',International Symposium on Cyber Security Cryptography and Machine Learning,Springer ,pp 231–248.

4. Ankit Kumar Jain, and B. B. Gupta,(2019) 'A machine learning based approach for phishing detection using hyperlinks information,Journal of Ambient Intelligence and Humanized Computing, Springer ,10(12),doi:10.1007/s12652-018-0798-z.

5. Noor Faisal Abedin, Rosemary Bawm, Tawsif Sarwar, Mohammed Saifuddin,Mohammd Azizur Rahman and Sohrab Hossain,(2020) 'Phishing Attack Detection using Machine Learning Classification Techniques,International Conference on Intelligent Sustainable Systems (ICISS), IEEE Explore ,doi: 10.1109/ICISS49785.2020

6.Ponnurangam Kumaraguru, LorrieFaith Cranor, and Laura Mather,(2009), AntiPhishing Landing Page: Turning a 404 into a TeachableMoment for End Users.Sixth Conference on Email and Anti-Spam ,pp- 16-17, Mountain View, California USA.

7.Chandrasekaran, M., Narayanan, M. and Upadhyaya, S. (2006) Phishing Email Detection Based on Structural Properties. Proceedings of 9th Annual NYS Cyber Security Conference, Albany, 14 June 2006, 2-8.

8.Dr.Radha Damodaram(2016). 'Study on phishing attacks and antiphishing tools'International Research Journal of Engineering and Technology (IRJET), Vol:03(01).

9.Kalaharshaa P.,and Mehtre B. M.,(2021) 'Detecting Phishing Sites - An Overview' arXiv:2103.12739v1 [cs.CR] 23 Mar 2021.

10. Edwin Donald Frauenstein,(2018) 'An Investigation Into Students Responses to Various Phishing Emails and Other Phishing-Related Behaviours'17th International Information Security South Africa Conference At: Pretoria, South Africa.

11.Shrusthi Patil, and Sudhir Dhage(2019);A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework,International Conference on Advanced Computing and Communication Systems (ICACCS),IEEE Explore,doi: 10.1109/ICACCS45965.2019.