Detection of phishing websites using Machine Learning

Varun (187R1A0504)

Uma Mahi(187R1A0520)

Vishanth(187R1A0531)

---***---

Abstract –

The risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest nowadays. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes.

Key Words: websites, URL detection, phishing

1.INTRODUCTION

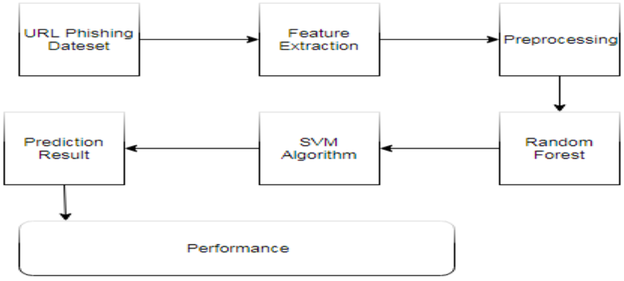
Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or another phishing site, or malware download.

The main feature of this project is that the system prevents users from visiting malicious websites by displaying a pop-up, it also displays the information regarding the website such as domain details and also gives suggestions for user on how to be safe from malicious URLS.

Defining the problem : Compromised URLs that are used for cyber-attacks are termed as malicious URLs. In fact, it was noted that close to one-third of all websites are potentially malicious in nature, demonstrating rampant use of malicious URLs to perpetrate cyber-crimes. A Malicious URL or a malicious web site hosts a variety of unsolicited content in the form of spam, phishing, or drive-by download in order to launch attacks.

2. Body of Paper

Phiishing is a fraudulent technique that uses social and technological tricks to steal customeridentification and financial credentials. Social media systems use spoofed e-mails from legiti-mate companies and agencies to enable users to use fake websites to divulge financial detailslike usernames and passwords. Hackers install malicious software on computers to stealcredentials, often using systems to intercept username and passwords of consumers' onlineaccounts. 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 thecontent in order to obtain their sensitive data. The primary objective of phishing is to gain cer-tain personal information for financial gain or use of identity theft. Phishing attacks are causing severe economic damage around the world. Moreover, Most phishing attacks targetfinancial/payment institutions and webmail, according to the Anti-Phishing Working Group(APWG) latest Phishing pattern studies [1].In order to receive confidential data, criminals develop unauthorized replicas of a real web-site and email, typically from a financial institution or other organization dealing with finan-cial data. This e-mail is rendered using a legitimate company's logos and slogans. Thedesign and structure of HTML allow copying of images or an entire website. Also, it is oneof the factors for the rapid growth of Internet as a communication medium, and enables themisuse of brands, trademarks and other company identifiers that customers rely on as authen-tication mechanisms. To trap users, Phisher sends "spooled" mails to as many people aspossible. When these e-mails are opened, the customers tend to be diverted from the legitimateentity to a spoofed website.



System Architecture

URL Phishing dataset : This dataset contains number of URL links for analysing their relevancy.

Feature Extraction : extracting the pre-defined features for a considered URL.

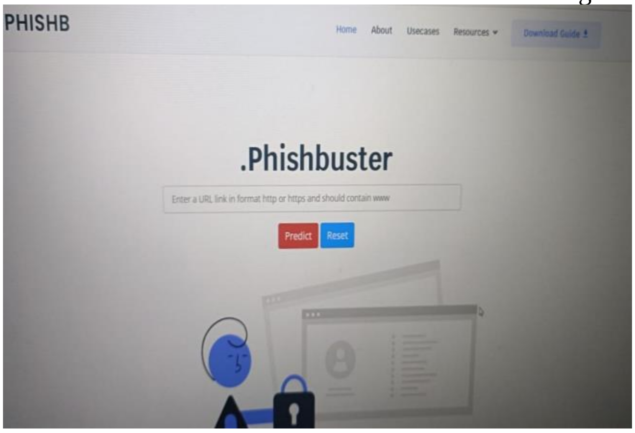
Preprocessing : Here, the relevant data is processed and the data that seems to be relevant and unrepeated is taken into the next step.

Random forest : Random forest algorithm is applied on the results that are obtained from decision tree.

Prediction Result : Here, after going through all the steps that are considered the result of our prediction is given to the user.

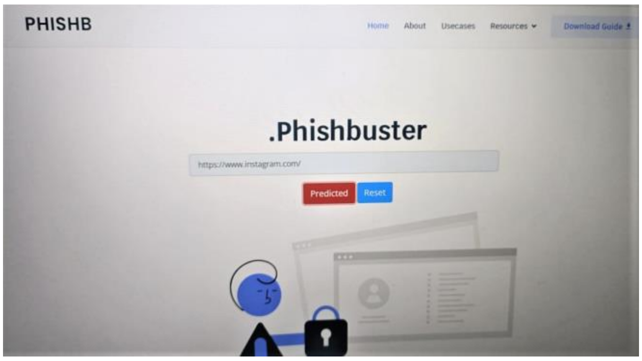
3.Working

a. Homepage : This home page indicates the user uploads the URL to check whether the URL is malicious or legitimate.



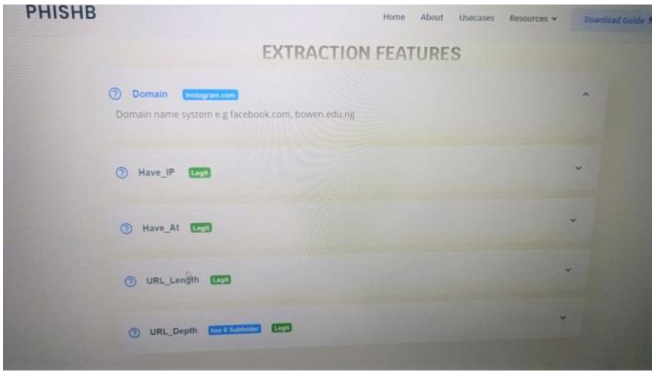
b.Malicious URL

When a user uploads the URL if it is malicious the detection system displays the result.



c.Domain Information

When the detection system detects the malicious URL and also displays the domain details of the URL.



EXTRACTION FEATURES

Domain instagram.com
Domain name system e.g facebook.com, bowen.edu.ng

Have_IP Legit

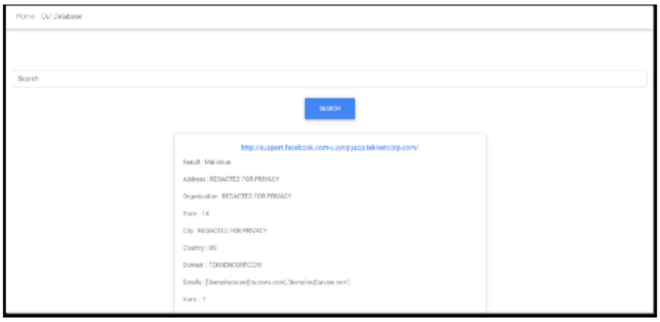
Have_At Legit

URL_Length Legit

URL_Depth has 0 Subfolder Legit

d.history

It contains information about searches made by the user.



http://support.facebook.com-uuyngiyacp.tekhencorp.com/

Result : Malicious

Address : REDACTED FOR PRIVACY

Organization : REDACTED FOR PRIVACY

State : TX

City : REDACTED FOR PRIVACY

Country : US

Domain : TEKHENCORP.COM

Emails : ['domainabuse@tucows.com', 'domains@arvixe.com']

Rank : 1

4.CONCLUSION

The detection model achieves the expected effect in experiments. However, considering that the network traffic in the test environment and the real network are different, and with the development of the Internet, types of malicious URL are more diverse. It is necessary to timely update the model in the actual scenario. Therefore, to better adapt to the requirements of various complex application scenarios, we plan to study how to simplify the detection model's architecture and shorten the training time while keeping the detection performance unchanged in the future.

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

We take this opportunity to express our profound gratitude and deep thanks to our guide for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark. We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) **Dr. M. Varaprasad Rao**, **Dr. G.Madhukar**, **Dr. T. S. Mastan Rao**, **Dr. Suwarna Gothane**, **Mr. A. Uday Kiran**, **Mr. A. Kiran Kumar**, **Mrs. G. Latha** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju**, Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged **to Dr. A. Raji Reddy**, Director for being cooperative throughout the course of this project
We also express our sincere gratitude to **Sri. Ch. Gopal Reddy**, Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help. Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

REFERENCES

1. Sadia Afroz and Rachel Greenstadt. 2011. Phishzoo: Detecting phishing websites by looking at them. In Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. IEEE.

2. A Astorino, A Chiarello, M Gaudioso, and A Piccolo. 2016. Malicious URL Detection via spherical classification. Neural Computing Applications (2016).