A Survey on Fishy URL Detection Using URL Features and CNN

Anudit Tribhuwan, Abhijit Taware, Hrushikesh Silam, Shubham Manale

(Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra.)

Abstract *– Phishing is a network type attack where the attacker creates the fake of an existing webpage to fool an online user into elicits personal Information. The prime objective of this review is to do literature survey on social engineering attack: Phishing attack and techniques to detect attack. NLP offers a natural solution for this problem as it is capable of analyzing the textual content to perform intelligent recognition and performing semantic analysis of text to detect malicious intent. In this type of cyber-attack, the attacker sends malicious links or attachments through phishing e-mails that can perform various functions, including capturing the login credentials or account information of the victim. This paper aims to provide a comprehensive and comparative study of various existing free service systems and research-based systems used for phishing website detection. The systems in this survey range from different detection techniques and tools used by many researchers.*

Keywords**: -** Mobile phones; phishing attack; security; anti-phishing

I. INTRODUCTION

Phishing basically uses social engineering techniques to trick users such as creating fake websites which clones with same attributes and design of the existing legitimate one. In a classic phishing attack a phisher send a link enclosed in a message to the user. The link redirects the user to the cloned malicious page which looks similar to the original webpage but is not and is intended to steal user's sensitive data. Such phishing attacks have proven to cause a lot of financial loss to various organizations.

Phishing is a fraudulent act that is used to deceive users over the Internet with the goal of obtaining their personal information. The attackers who plan phishing attacks are commonly termed as phishers. Phishing became a serious cyber threat in 1996 when phishers stole the user names and passwords of AOL users. In most cases, a successful phishing attack is accomplished using email spoofing and website spoofing techniques as shown in following figure

Phished Email
Phished Website
Successful Phishing

Figure: - Successful phishing: A combination of phished email and websites.

In this survey, we review the phishing website detection systems which use advanced tools and techniques to provide promising results in this domain. We specifically focused on the work which presented the feature representation model with an advanced machine learning algorithm for development.

II. BACKGROUND THEORY

Phishing is the fraudulent attempt to obtain sensitive information such as usernames, passwords and credit card details, often for malicious reasons, by disguising as a trustworthy entity in an electronic communication. Phishing attack can be implemented in various form like Email phishing, Website phishing, spear phishing, Whaling, Tab napping, Evil twin phishing etc. To avoid this phishing attack various anti-phishing solutions should be use. There are various anti phishing solutions such as Blacklist, heuristic, visual similarity, machine learning etc.

PHISHING ATTACKS

In a phishing attack, attackers can use social engineering and other public information resources, including social networks like LinkedIn, Facebook and Twitter, to gather background information about the victim's personal and work history, interests and activities. With this pre-discovery, attackers can identify potential victims' names, job titles and email addresses, information about the names of key employees in their colleagues and organizations.

The common information that is stolen by a phishing attack is listed as follows:

- User account number
- User passwords and user name
- Credit card information
- Internet banking information

III. LITERATURE REVIEW

A. Lightweight phish detector (LPD)

Varshney et al. focused on the need of lightweight phishing detection approach using search engines. Authors identified the lightest possible features (page title and domain name) that can be extracted from a webpage without a complete webpage loading. Based on this, authors developed an intelligent anti-phishing chrome extension named lightweight phish detector (LPD). LPD not only detects but also suggests the authentic webpage to the user when a user reaches a deceptive or phishing page on the browser.

B. C4.5 decision tree algorithm

This paper proposes a efficient way to detect phishing URL websites by using c4.5 decision tree approach. This technique extracts features from the sites and calculates heuristic values. These values were given to the c4.5 decision tree algorithm to determine whether the site is phishing or not. Dataset is collected from Phish Tank and Google. This process includes two phases namely pre-processing phase and detection phase. In which features are extracted based on rules in pre-processing phase and the features and their respected values were inputted to the c4.5 algorithm.

Accuracy: - c4.5 algorithm obtained 89.40% accuracy.

C. Machine Learning Anti- Phishing System (MLAPT)

This paper proposed a system that determines phishing mails using two existing systems, Machine Learning Anti- Phishing System (MLAPT) and Phish zoo. The Phish zoo system uses the visually based approach for phishing detection while the Machine Learning Anti- Phishing System (MLAPT) helps in determining the mails present on the system into a phishing or benign category. The presented model proved effective to manage personal sensitive information on social networking websites.

D. Authentication technique to reduces phishing attack

In this proposed system the author used steganography approach to hide our profile. The password strength should not be weak. This methodology is that the user password may be an image that is the authentication process to identified user.

Advantage: - It is more secure technique to hide our password from the attacker.

Limitations: - For password securing no proper formwork is suggested in social engineering.

E. LSTM:

Minh Nguyen, Toan Nguyen, Thien Huu Nguyen presented a framework with hierarchical long short-term memory networks (HLSTMs) and attention mechanisms to model the emails simultaneously at the word and the sentence level. Expectation is to produce an effective model for anti-phishing and demonstrate the effectiveness of deep learning for problems in cybersecurity. The recall, F1- score and precision are used to evaluate the performance of the models for detecting phishing emails are compared with the SVM baselines in two different settings when the email headers are not considered. 2 types of data: without header and with header.

Accuracy: - Without header accuracy of 98.1% and with header accuracy of 99%.

F. Anti-phishing single sign on model using QR Code: -

This technique addresses the problem of phishing on single sign on authentication. **Single sign on** is an authentication process that permits users single username and password to access multiple applications or websites.

The technique uses QR codes since they do not need mobile network data to read the data and it can store a large amount of information. There are two phases in this approach;

- **User Registration Phase: -**

In User Registration the user receives a secret key which is later used in verification phase to get access to the requested service.

- **User Verification Phase:-**

In verification phase user requests service from the service provider which sends the user identity to the identity provider.

IV. PROPOSED SYSTEM

First of all, the hacker has to create a phishing website which looks legitimate in order to lure the victim. Then, host the website on the internet. If a victim visits the phishing website, it convinces him/her to enter some confidential information. The hacker then acquires the entered data which can be misused.

We aim to use WhoIs features of URL as the basis of detecting phishing websites. We propose a novel solution, Phishing Detection using Soft Computing and Machine Learning, to efficiently detect phishing web pages using URL and WhoIs features. The Convolution Neural Network is used to train the network and detect whether the website is legitimate or not.

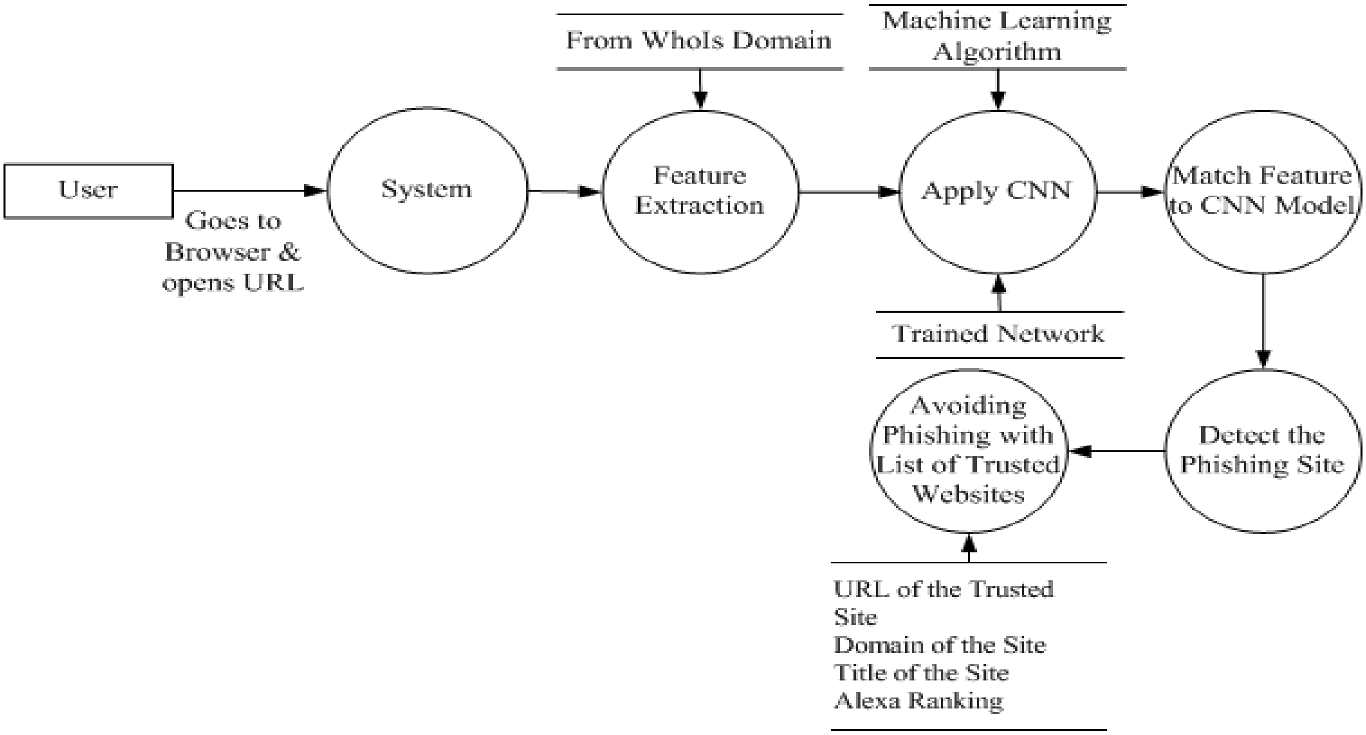

Figure: - Flow chart of Proposed System.

V. CONCLUSION

Phishing URL detection plays a pivotal role for many cyber security software and applications. In this paper, we researched and reviewed works based on the advanced machine learning techniques and approaches that promise a fresh approach in this domain. This article includes summary of the reviewed works after a systematic and comprehensive study on Phishing Website Detection systems.

We believe that the presented survey would help researchers and developers with the insight of the progress achieved in the past years. Despite the tremendous progress in the field of cyber security, phishing website detection still poses a challenging problem with the ever evolving technology and techniques. In the proposed technique, the system model is built to detect phishing sites by using some neural network algorithms like Convolutional Neural Network (CNN).

REFERENCES

[1] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[2] Neda Abdelhamid, Fadi Thabtah, Hussein Abdel-jaber “Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features” IEEE 2017.

[3] Longfei Wu, Xiaojiang Du, and Jie Wu “MobiFish: A Lightweight Anti-Phishing Scheme for Mobile Phones” IEEE 2014.

[4] LongfeiWu, Xiaojiang Du, and Jie Wu, “Effective Defense Schemes for Phishing Attacks on Mobile Computing Platforms” IEEE 2015.

[5] Guang-Gang Geng, Zhi-Wei Yan, Yu Zeng and Xiao-Bo Jin “RRPhish- Anti-Phishing via Mining Brand Resources Request” 2018 IEEE International Conference on Consumer Electronics (ICCE)

[6] Sadia Afroz and Rachel Greenstadt “PhishZoo: Detecting Phishing Websites By Looking at Them” IEEE 2011.

[7] Muhammet Baykara and Zahit Ziya Gürel “Detection of phishing attacks” IEEE 2018

[8] Mohammed Nazim Feroz,Susan Mengel "Phishing URL detection using URL Ranking" International Congress on Big Data 2015 IEEE.

[9] Luong Anh Tuan Nguyen†, Ba Lam To†, Huu Khuong Nguyen† and Minh Hoang Nguyen* † Faculty of Information Technology "Detecting Phishing Web sites: A Heuristic URL-Based Approach" International Conference on Advanced Technologies for Communications 2013.

[10] Ji Hua 1,2, Zhang Huaxiang 1,2 "Analysis on the Content Features and Their Correlation of Web Pages for Spam Detection" IEEE 2015.

[11] Samuel Marchal, Jérôme François, Radu State, and Thomas Engel "PhishStorm: Detecting Phishing With Streaming Analytics" IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, 2014.

[12] Luong Anh Tuan Nguyen1, Ba Lam To2, Huu Khuong Nguyen1 and Minh Hoang Nguyen31 Faculty of Information Technology "A Novel Approach for Phishing Detection Using URL-Based Heuristic" IEEE 2014.

[13] Jian Mao1,2, Pei Li 1, Kun Li1, Tao Wei3, and Zhenkai Liang4 "BaitAlarm: Detecting Phishing Sites Using Similarity in Fundamental Visual Features" 5th International Conference on Intelligent Networking and Collaborative Systems 2013.