Smarteye - Fishy URL Detection Using URL Features and CNN: SURVEY

Prof. Sujay Pawar[1], Aishwarya Baynoor[2], Amit Kumar[3], Arnav Das[4], Yash Kulkarni[5]

[1](Professor, Department of Computer Engineering, Dr.D.Y.Patil Institute of Engineering and Technology, Pune, Maharashtra, India)

[2345](Students, Department of Computer Engineering, Dr.D.Y.Patil Institute of Engineering and Technology, Pune, Maharashtra, India)

Abstract *– Phishing URL is a widely used and common technique for cyber security attacks. Phishing is a cybercrime that tries to trick the targeted users into exposing their private and sensitive information to the attacker. The motive of the attacker is to gain access to personal information such as usernames, login credentials, passwords, financial account details, social networking data, and personal addresses. These private credentials are then often used for malicious activities such as identity theft, notoriety, financial gain, reputation damage, and many more illegal activities. 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

The advancement of internet is resulting in attracting more and more users into this huge Internet Sea. There are a lot of perks of using internet, one can buy stuff online, way of learning and gaining knowledge has improved, etc. On the contrary, possible threats come hand in hand. One of them is Phishing Attack. Phishing is an attack where a legitimate user is deceived to disclose sensitive information and assets with economic value. Loss of such sensitive information might cause potential economic or reputational harm an organization.

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.

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.

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

II. LITERATURE REVIEW

A. Random Forest Method

The author has introduced a model with answer for recognize phishing sites by utilizing URL identification strategy utilizing Random Forest algorithm. Show has three stages, namely Parsing, Heuristic Classification of data, Performance Analysis. Parsing is used to analyze feature set. Dataset gathered from Phish tank. Out of 31 features only 8 features are considered for parsing.

Accuracy: - Random forest method obtained accuracy level of 95%.

B. Convolution Neural Network (CNN) And Long Short-Term Memory (CNN-LSTM)

In this paper authors [7] made a comparative study to detect malicious URL with classical machine learning technique like; logistic regression using bigram, deep learning techniques like convolution neural network (CNN) and CNN long short-term memory (CNN-LSTM) as architecture. The dataset collected from Phish tank, Open Phish for phishing URLs and dataset Malware Domainlist, Malware Domains were collected for malicious URLs. In this paper authors used Tensor Flow in conjunction with Keras [for deep learning architecture.

Accuracy: - As a result of comparison, CNN-LSTM obtained 98% 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.

III. PROPOSED SYSTEM

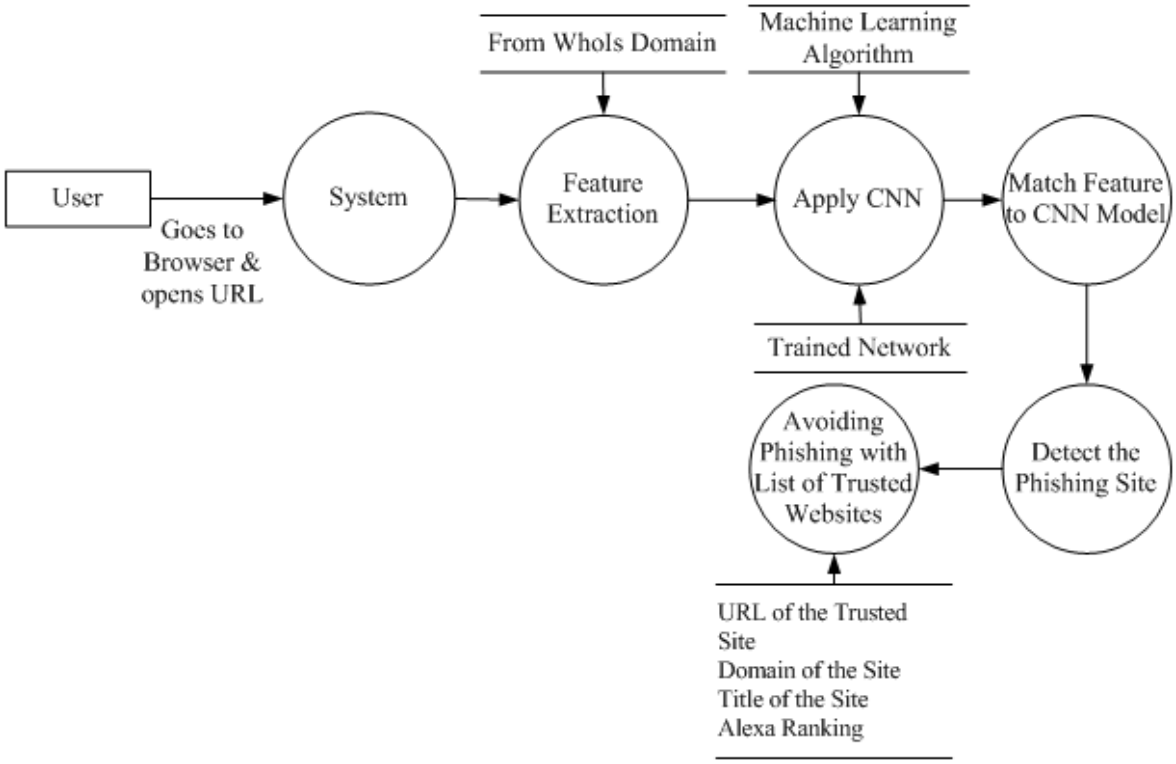

Figure: - Flow chart of Proposed System.

First of all, the phisher has to create a phishing website to lure the victim which seems as legitimate one. Then, host the site on the internet for use of victim secrete information. If victim visit phishing website, it convinces the victim to enter some confidential information. Phisher then acquire some entered data and later it can be misuse by phisher.

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 finally detect the site is Phishing or not.

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.