Implementing Smarteye - Fishy URL Detection Using URL Features and CNN

Vaibhav Gaikwad, Ankur Bakre, Sourish Joshi, Satyam Patil, Prof. Yogita Narule

Department of Information Technology , I2IT Pune, Maharashtra.

Abstract *– Phishing attacks are very common but are the least defended security menaces today. An approach is proposed which makes use of Natural Language Processing techniques to detect statements which are indication of phishing attacks. 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. The standard way to specify page layouts is through the style sheet (CSS), the developed algorithm detects similarities in key elements related to CSS. Phishing detection includes approach that uses profiles of trusted websites' appearances to detect phishing.*

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

I. INTRODUCTION

Phishing is defined as the fraudulent acquisition of confidential data by the intended recipients and the misuse of such data. **The phishing attack is often done by email.** An example of Phishing; as if e-mail appear to be from known web sites, from a user's bank, credit card company, e-mail, or Internet service provider. Generally, personal information such as credit card number or password is asked to update accounts.

These emails contain a URL link that directs users to another website. This site is actually a fake or modified website. When users go to this site, they are asked to enter personal information to be forwarded to the phishing attacker.

PHISHING ATTACKS

The **aim** is to steal sensitive data such as credit card and login information or to install malicious software on the victim's machine. Phishing is a common type of cyber-attack that everyone must learn to protect them. Phishing is start with a fake e-mail or other type of transmission designed to attract a victim. In this type of attack, the message appears to come from a trusted source.

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.

Phishing is also used to learn someone's password or credit card information. With the help of e-mail prepared as if coming from a bank or official institution, computer users are directed to fake sites.

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. Support Vector Machine:

Fergus Toolan and Joe Carthy [1] provide the instances that are very small consisting of only five features. Using over 8,000 emails, approximately half of which were phishing emails and the remainder legitimate, results of evaluation are presented. Adwan Yasin et al [2] proposed a model that applied the knowledge discovery procedures making use of five popular classification algorithms among which SVM was one and achieved a notable enhancement in classification accuracy. Srishti

Rawal et al [3] Aim is to use the least number of features to develop a system which provides higher accuracy and study the variation of features. The features where extracted using regular expressions and NLTK.

Accuracy: - Maximum accuracy of 99. 87% is achieved in classification of emails using SVM.

B. Authentication technique to reduces phishing attack

- **Methodology: -** In this we used steganography techniques 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.
- **Advantages: -** It is more secure technique to hide our password from the attacker.
- **Limitations: -** For password securing no proper formwork is suggested in social engineering.

C. Second Level Domain (SLD):-

The login form is detected then the tool extracts **second level domain (SLD)** name from the URL. The SLD is mapped with its brand name using the mapping list that is generated. Screen shot of webpage is captured and converted to text using OCR tool.

If there are some typical sensitive terms in the text then the user will get the warning. If SLD is not present in the text that is extracted from the screenshot then it is possibly a phishing web page.

D. Naive Bayes:

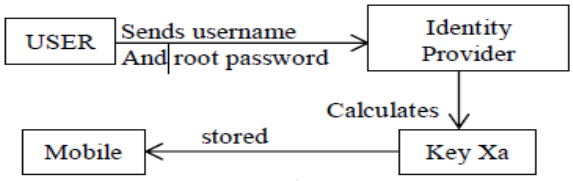
Adwan Yasin et. al [2] introduces the concept of phishing terms weighting which evaluates the weight of phishing terms in each email. The pre-processing phase is enhanced by applying stemming and WordNet. Authors in [4] discussed a technique from which success rate of 89% has been achieved against phishing attacks coming from email messages. Srishti Rawal and co-authors discussed a model in [3] where 9 features were extracted from all emails in a self-made dataset which consists of n phished emails and m ham emails. These features are given to the classifiers and results noted.

E. 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.

F. Random Forest:

Andronicus A et. al proposed a classifier with better prediction accuracy and fewer numbers of features. A set of prominent phishing email features were extracted from a dataset consisting containing 2000 phishing and ham emails. [5] S. Rawal et. al discussed phished email classifier in which 9 features were extracted from all emails in a self-made dataset which consists of n phished emails and m ham emails[3].

Accuracy: - The machine learning algorithm resulted in classification accuracy of 99.7%

III. PROPOSED SYSTEM

Phishing is a criminal scheme to steal the user's personal data and other credential information. It is a fraud that acquires victim's confidential information such as password, bank account detail, credit card number, financial username and password etc. and later it can be misuse by attacker. We aim to use fundamental visual features of a web page's appearance as the basis of detecting page similarities. We propose a novel solution, Phishing Detection using Soft Computing and Machine Learning, to efficiently detect phishing web pages using URL and CSS features.

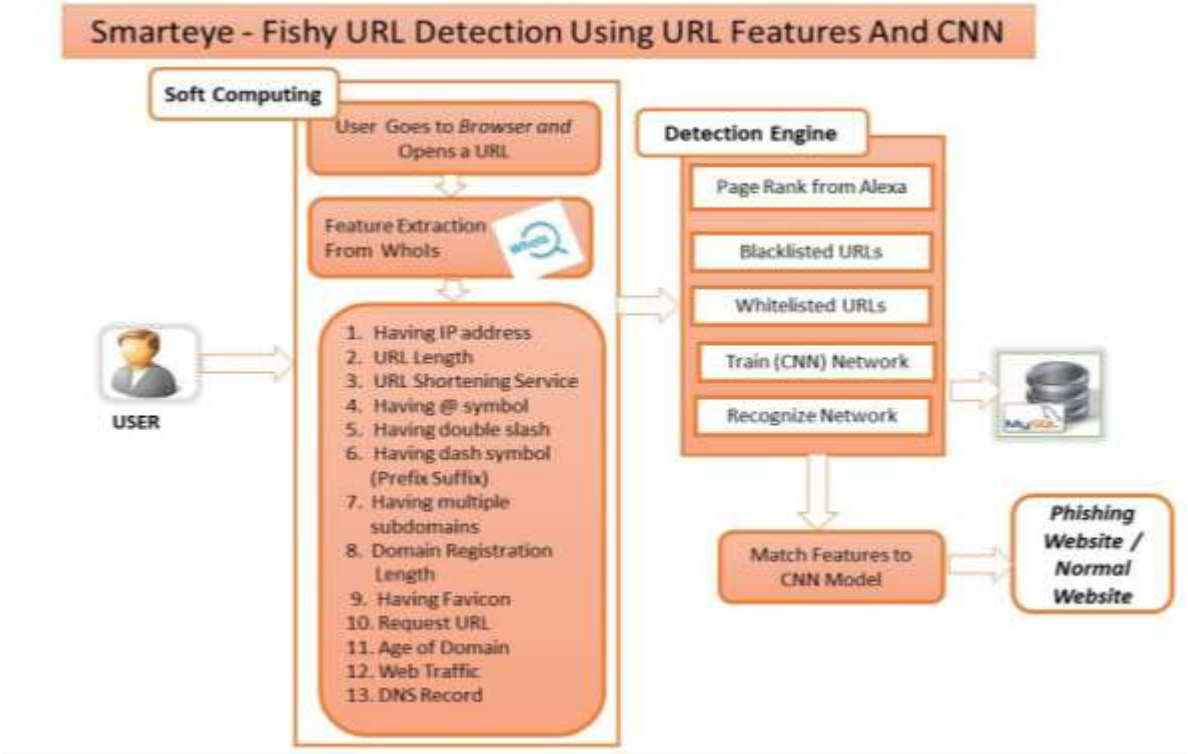


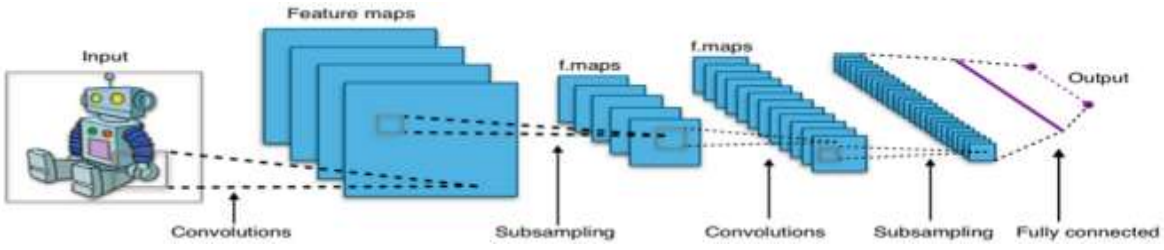
Figure: - System Architecture.

We will be using some algorithms like CSS Detection Algorithm, ObURL Detection Algorithm and Feed Forward Neural Network for phishing detection.

IV. ALGORITHMS USED

Convolution Neural Network Traditional feature learning methods rely on semantic labels of images as supervision. They usually assume that the tags are evenly exclusive and thus do not pointing out towards the complication of labels. The learned features endow explicit semantic relations with words. We also develop a novel cross-modal feature that can both represent visual and textual contents. CNN is a method of categorizing the images as a part of deep learning. In which we apply a single neural network to the full image. The steps in CNN are as follows: convolution, subsampling, activation and full connectedness.

Step 1: Convolution it is the primary layers that accept an input signal are called convolution filters. Convolution is a procedure where the network tries to tag the input signal by referring to what it has learned in the past.



Step 2: Subsampling Inputs from the convolution layer can be smoothened to decrease the sensitivity of the filters to noise and variations. This smoothing procedure is labeled as sub- sampling, and can be attained by taking averages or considering the maximum over a sample of the signal.

Step 3: Activation the activation layer manages the signal flows from one layer to the subsequent Output signals which are strongly connected with past references would activate more neurons, enabling signals to be propagated more efficiently for identification.

Step 4: Fully connected the final layers in the network are fully connected, such that the neurons of preceding layers are connected to every neuron in subsequent layers. This imitates high Level reasoning where all feasible path ways from the input to output are measured.

V. RESULTS AND DISCUSSIONS

Normal Sit: - Non Phishing Site.

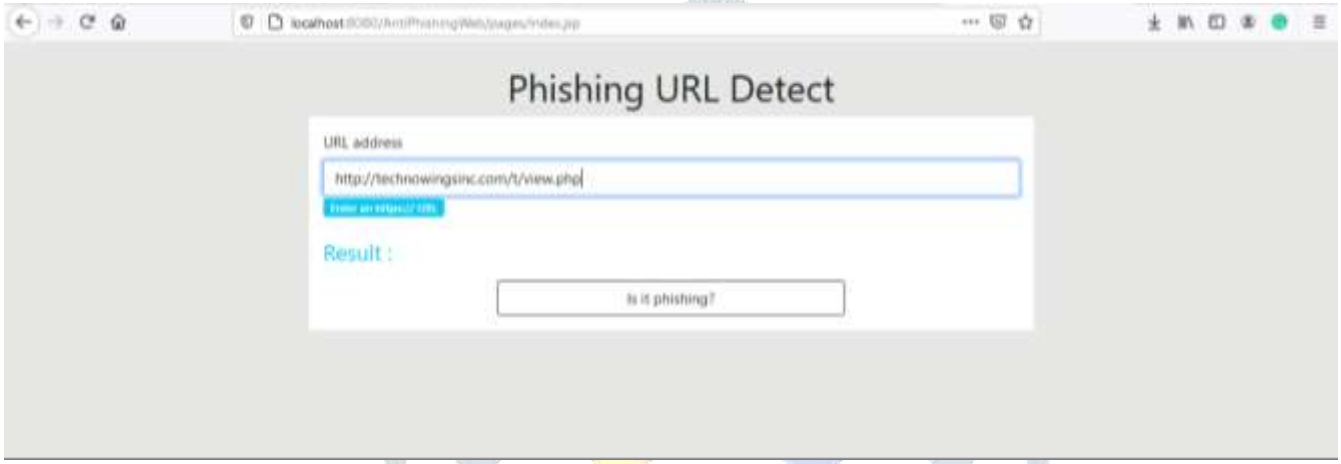


Figure: - Input to check the URL address.

Figure: - The result- Non phishing (Normal Site)

Abnormal Sit: - Phishing Site.

Figure: - Input to check the URL address.

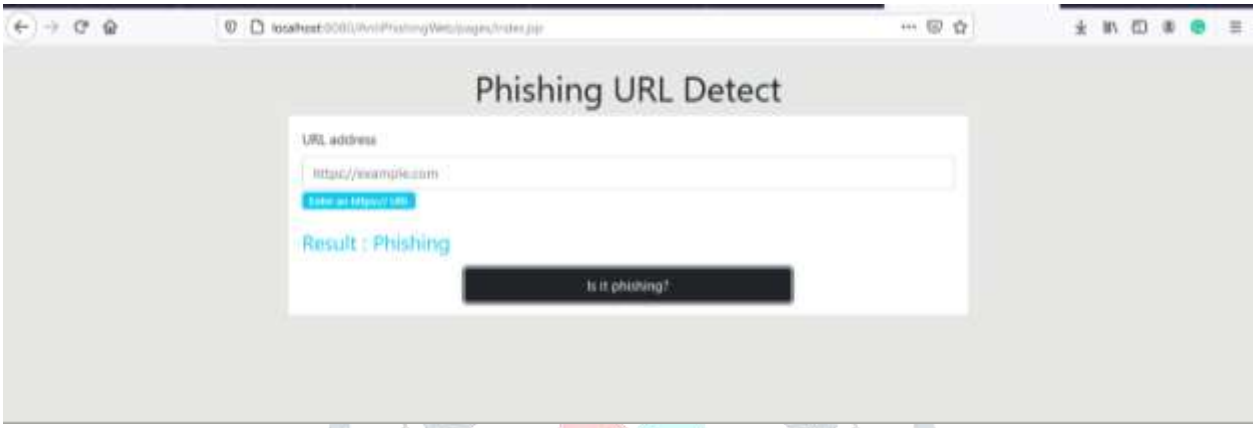


Figure: - The result- Phishing (Abnormal Site)

CONCLUSION

Phishing is a criminal scheme to steal the user's personal data and other credential information. It is a fraud that acquires victim's confidential information such as password, bank account detail, credit card number, financial username and password etc. and later it can be misuse by attacker.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.

REFERENCES

[1] Fergus Toolan, Joe Carthy, "Phishing Detection using Classifier Ensembles," 2009 eCrime Researchers Summit, Tacoma, WA, USA,2009.

[2] 2. Andronicus A. Akinyelu, Aderemi O. Adewumi, "Classification of Phishing Email using Random Forest Machine Learning",Vol.2014, Article ID 425731, Hindawi Publishing Corporation, April 2014.

[3] Adwan Yasin, Abdel Munem Abuhasan, "An Intelligent Classification Model for Phishing Email Detection", Vol.8, No.4, International Journal of Network Security & Its Applications (IJNSA), July 2016.

[4] Elif Yerli, Ibrahim Sogukpinar, "Email Phishing Detection and Prevention by using Data Mining Techniques", 2017.

[5] Srishti Rawal, Bhuvan Rawal, Aakhila Shaheen, Shubham Malik, "Phishing Detection in Emails using Machine Learning," Vol.12 –No. 7, International Journal of Applied Information Systems (IJAIS), October 2017. 42

[6] S. H. Gunawardena, D. Kulkarni and B. Gnanasekaraiyer, "A steganography-based Framework to Prevent Active Attacks during User Authentication," 8thInternational Conference on Computer science & Education (ICCSE), 2013, pp. 383 - 388.

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

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

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

[10] 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)

[11] 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.

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

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

[14] 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.

[15] 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.