An Implementation of Fraudulent URL Detection with Advanced Feature and ML Technique

Rajesh Agrawal
(Technical Architect, Technowings Pune)

Sourabh Kumbhar
(Research Intern, Technowings Pune)

Abstract – "Phishing is a fraudulent attempt, usually made through email, to steal your personal information". Phishing is becoming more malicious day by day and its detection is very important. It is one of the social engineering methods that gathers personal information through websites such as malicious websites and deceptive e-mail to canvass personal information from a company or an individual by prance as a trustworthy entity or organization. Phishing often attacks email by using as a vehicle and even sending messages by email to users that represent a part of a company or an institution who perform business such as financial institution, banking etc. We reviewed various data mining algorithms for evaluation of the features in order to get a better understanding of the structure of URLs that spread phishing. The fine-tuned parameters are useful in selecting the appropriate machine learning algorithm for separating the Phishing sites from benign sites.

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

I. INTRODUCTION

What is Phishing?

First of all, the phisher has to create a phishing website to lure the victim who seems as legitimate one [2]. 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.

Host based analysis

Host-based features explain "where" phishing sites are hosted, "who" they are managed by, and "how" they are administered. [14]We use these features because phishing Web sites may be hosted in less reputable hosting centers, on machines that are not usual Web hosts, or through not so reputable registrars.

The block schematic for the host based analysis is shown in following Figure.

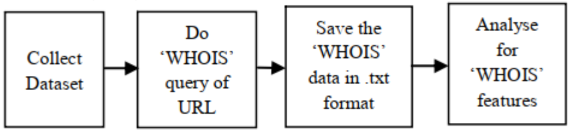

Collect Dataset
Do 'WHOIS' query of URL
Save the 'WHOIS' data in .txt format
Analyse for 'WHOIS' features

Figure: - Block diagram for host based analysis [14]

WHOIS

WHOIS Features: WHOIS [12] properties gives details about the date of registration, update and expiry, who is the registrar and the registrant. If phishing sites are taken down frequently, the registration dates will be newer than for legitimate sites. A large number of phishing websites contain IP address in their hostname [13]. So getting the details of such hostnames will be helpful in efforts to point to phishing sites, which can be obtained from the Whois features.

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 [13]. The convolution Neural Network is used to train the network and finally detect the site is Phishing or not. 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 [14]. 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 [6, 9]. 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

<u>AVOIDING PHISHING ATTACKS</u>

A whitelist in the context of phishing detection is simply a list of trusted websites. For CSS detection to work properly, the list contains more than just the URL of the trusted website. Each entry in the whitelist database contains six strings: the URL of the trusted site, the domain of the site, the title of the site, the CSS filename, the CSS domain, and the CSS content of the file.

a. The URL of the trusted site:

The URL of the trusted site is used to periodically update the CSS information in the database. This is the URL of the site such as "https:\\signin.ebay.com".

b. The domain of the site:

The domain of the trusted site is the domain of the URL such as "signin.ebay.com" and is used to determine whether the current page displayed in the browser is on the whitelist or not.

c. The CSS filename:

The CSS filename is the filename of the CSS file such as "paypal.css" and can also be used during CSS content detection to speed up detection by matching potential phishing site CSS filenames with filenames in the whitelist database.

d. The CSS domain:
The CSS domain is the domain of the location of the CSS file such as "secureinclude.ebaystatic.com". Often the domain is the same as the site domain, but in other cases such as eBay, the CSS file is hosted on a different domain. Storing the CSS domain is essential because if a match is found on a website not in the whitelist, then it is most likely a phishing site linking to the actual CSS file location of the legitimate site.

e. The CSS content of the file:
The CSS content is the actual text contained in the CSS file that contains all of the style information. The CSS content is used to compare with the CSS content of a possible phishing site in order to determine if there is a match with a legitimate site.

II. LITERATURE REVIEW

A. Non-content based approaches

Now a days phishing detection schemes can be usually divided into two categories:

Heuristics-based schemes:-
Heuristics-based schemes largely depend on features extracted from URL and HTML source code, and then other techniques such as machine learning are used to determine the validity.

Blacklist based schemes:-
Blacklist-based schemes can only detect phishing sites that are in the blacklist but cannot detect zero-day phishing attacks that have appeared for days or even hours. It is possible that new phishing sites may have already stolen user credentials or even expired before being added into the blacklist.

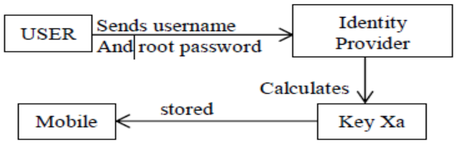
B. Detection of phishing URL using Artificial Neural Network

It is a method to classify the Uniform Resource Locator (URL) into Phishing URL or Non -phishing URL is designed. To improve the performance of ANN, use of particle swam optimization and classification training should be done. A dynamic approach for detecting phishing techniques is proposed which uses a single layer artificial neural network. In this paper, In the First step of the technique value of six heuristics are calculated using this algorithm.

C. 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.

D. Earth's Mover Distance (EMD)

EMD is a metric of the similarity between two probability distributions over a region. The closer two images are, the smaller the EMD value will be about using the hash value of images in their work [19]. The authors have also calculated pHash of a screenshot and evaluate the difference using hamming distance between two hash values. They present experimental results illustrating that adding a small change to the original image will also lead a small increase in the hamming distance.

E. RRPHISH

Because of maximum concurrent connections limitation to the same domain for browsers, brand sites usually run resource content on another domain, such as PayPal runs CSS, JS and image files on paypalobjects.com.
RRPhish can automatically extend the blacklist which will be an effective complement to the blacklisting method. For different application scenarios, contain different algorithms with different complexity, such as heuristic rules or machine learning algorithms.

III. SYSTEM ARCHITECTURE

As we know, the attacker uses the number of methods to obfuscate the URL. So, it is complex to detect all that attacks but the ObURL detection algorithm can detect the maximum number of URL obfuscation phishing attacks because following test cases are perform for checking the phishing site emails.

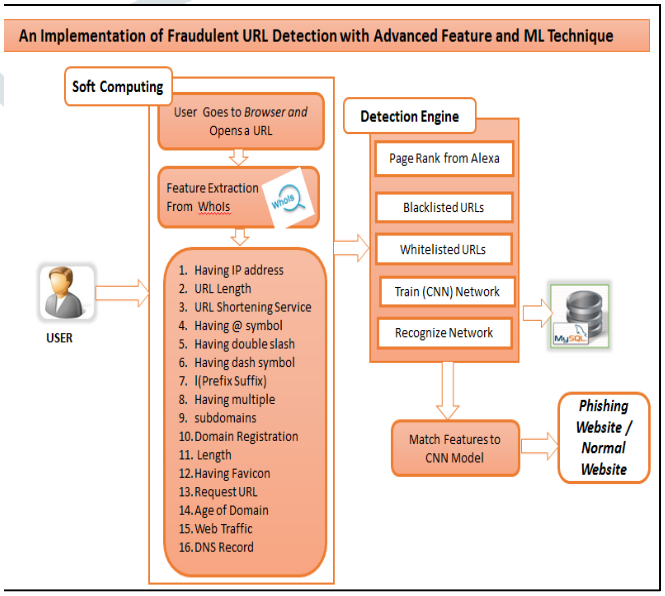


Figure: - System Architecture

1. Who is Domain

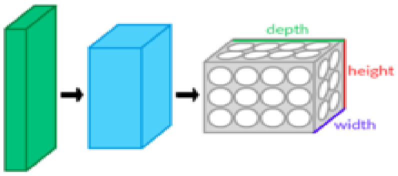
WHOIS is a question and reaction convention that is broadly utilized for questioning databases that store the enlisted clients or trustees of an Internet asset, for example, an area name, an IP address square or a self-governing framework, but on the other hand is utilized for a more extensive scope of other data. The convention stores and conveys database content in a comprehensible format. A WHOIS is a way for you to search the public database for information about a specific domain, such as the expiration date, current registrar, registrant information, etc.

TABLE: - The following features are extracted from Whois from URLs

Features	Significance
Having IP Addr	If IP address is used in domain name, then website is phishing.
URL_Length	Legitimate URLs have length of nearly 75 characters, URL length more than 75 is Phishing sites.
Shortening_Service	Link sharpeners are used to fool people.
Having_At_Symbol	Websites having an @ symbol are Phishy in
Double_slash_redirecting	If there is '/1' then it can be categorized as a Phishing Website.
Having_Sub_Domain	Legitimate Websites use only domain generally upto two level. Websites having more than 3 dots do it to include more domains within a domain, are generally Phishy.
URL_of_Anchor	In legitimate websites the anchor tag is connected to the same domain as the source code, Phishy websites have different domains.
Links_in_tags	Links in tags lead to some fraudulent websites.
Abnormal_URL	This feature is extracted from Who is Database, Legitimate websites' main identity is in the URL
Age_of_domain	Legitimate websites have an age of six months; websites with more than this age can be classified as Phishing.
Page_Rank	Phishing websites will have low page rank due to lack of links pointing to them.
Links_Pointing_to_page	Phishing websites have links pointing to zip files that automatically get downloaded containing malware.
Favicon	Many existing user agents such as graphical browsers and newsreaders show favicon as a visual reminder of the website identity in the address bar websites.
DNS (Domain Name System) Record	If the DNS record is empty or not found then the website is classified as "Phishing", otherwise it is classified as "Legitimate.
Web Traffic	Web traffic is the amount of data sent and received by visitors to a website. Phishing websites will create huge web traffic.
Website Traffic	This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit.

2. Built Detection Model using Convolution Neural Network

The system can detect the phishing site using Convolution Neural Network (CNN) technique. A CNN consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of Convolutional layers, pooling layers, fully connected layers and normalization layers. CNN will be used to train the data analytics engine for recognizing the phishing site URL.



3. Avoiding phishing attacks:

A whitelist in the context of phishing detection is simply a list of trusted websites.

1. The URL of the trusted site:

The URL of the trusted site is used to periodically update the information in the database. This is the URL of the site such as "https:\\signin.ebay.com".

2. The domain of the site:

The domain of the trusted site is the domain of the URL such as "signin.ebay.com" and is used to determine whether the current page displayed in the browser is on the whitelist or not.

3. The title of the site:

The title of the trusted site is the page title of the site such as "Welcome to eBay" and can be used to speed up the matching potential of phishing site titles with titles in the whitelist Database.

4. Alexa Ranking

In case your site is ranked relative to other sites, changes in traffic to other sites affect your site's rank. Every day, Alexa estimates the average daily visitors and page views to every site over the past 3 months. The site with the highest combination of visitors and page views over the past 3 months is ranked #1. As phishing websites live for a short period of time, they may not be recognized by the Alexa database (Alexa the Web Information Company., 1996). if the domain has no traffic or is not recognized by the Alexa database, it is classified as "Phishing". Otherwise, it is classified as "Suspicious".

IV. ALGORITHMS USED

Convolutional Neural Network Algorithm

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.

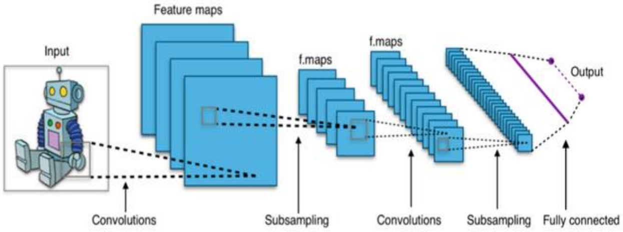


Figure: - CNN Algorithm

- **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. EXPERIMENTAL RESULTS

There are following outcomes of our proposed system;

➢ **Dataset**

- **Total No of Instances→ 11055**
- **Phishing→ 6157**
- **Non phishing→ 4898**

➢ **Result and Analysis**

We analyzed the prepared URL feature dataset using Naïve Bayes, Random Forest and SVM classifying algorithms in WEKA.

The performance is then evaluated based on Confusion matrix, Detection Accuracy, True Positive Rate and False Positive Rate. The result is tabulated in TABLE 1. The Support Vector Machine has the highest Success Rate compared to other selected classifying algorithms in WEKA.

TABLE: - Classifier Performance – WEKA

Classifier	Confusion Matrix	Accuracy (%)	Error Rate (%)
Naïve Bayes	**a b** 3471 1427 2317 3840	66.133%	33.867%
SVM	**a b** 4006 892 1805 4352	75.6038 %	24.3962%
Random Forest	**a b** 3605 1293 1322 4835	76.3455 %	23.6545%
CNN	**a b** 3682 1216 1196 4961	78.1818 %	21.8182%

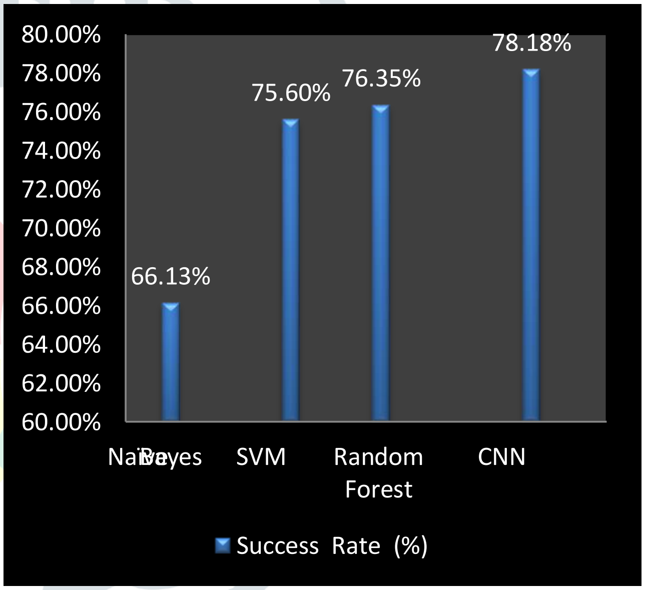


Figure: - The Comparison of the classifiers

The above figure shows a comparison of confusion matrix, Detection Accuracy and Error Rate of SVM, Naïve Bayes, Random Forest and CNN classifiers.

VII. CONCLUSION

Several features are compared using various data mining algorithms. The results points to the efficiency that can be achieved using the Who is features. To protect end users from visiting these sites, we can try to identify phishing URLs by analyzing their host-based features. A particular challenge in this domain is that criminals are constantly making new strategies to counter our defense measures. To succeed in this contest, we need algorithms that continually adapt to new examples and features of phishing URLs.

The aim of the proposed system is 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] Neda Abdelhamid, Fadi Thabtah, Hussein Abdel-jaber “Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features” IEEE 2017.

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

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

[4] 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)

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

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

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

[8] 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.

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

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

[11] 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.

[12] 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.

[13] Garera S., Provos N., Chew M., Rubin A. D., “A Frameworkfor Detection and measurement of phishing attacks”, In Proceedings of the ACM Workshop on Rapid Malcode (WORM), Alexandria, VA.

[14] Joby James, Sandhya L, Ciza Thomas “Detection of Phishing URLs Using Machine Learning Techniques” 2013 International Conference on Control Communication and Computing (ICCC).

[15] A. MahaLakshmi, N. Swapna Goud, Dr. G. Vishnu Murthy “A Survey on Phishing And It’s Detection Techniques Based on Support Vector Method (SVM) and Software Defined Networking(SDN)” International Journal of Engineering and Advanced Technology (IJEAT)

[16] Archit Shukla 1, Lalit Gehlod 2 “A survey on phishing detection and prevention technique” International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 3 Issue 5 may, 2014