RESEARCH ARTICLE OPEN ACCESS

Classification And Detection Of Phishing Websites

P.Hema Sujatha [1], S.Sushma Sree [2], N. Vinay Sreenath [3], S. Suresh [4],
Dr.Bala Brahmeswara Kadaru [5]
[1],[2],[3],[4] Dept.of CSE, Gudlavalleru Engineering College - Gudlavalleru,
[5] Assistant Professor, Dept.of CSE, Gudlavalleru Engineering College - Gudlavalleru,

ABSTRACT
Phishing is described as the art of imitating a website of a creditable firm intending to grab user's private information such as usernames, passwords, sensitive information and social security number. Phishers use the websites which are visually and conceptually similar to those real websites. As technology continues to grow, phishing techniques started to improve their progress rapidly and this needs to be prevented by using anti-phishing mechanisms/practices to detect phishing. Machine learning is a powerful tool used to achieve success against phishing attacks. This paper surveys the features used for detection and also the detection techniques using machine learning. The classifiers were tested with a data set containing 2,017 real world URLs where each could be categorized and classified as a legitimate site or phishing site. The results of the experiments show that the classifiers we considered were successful in distinguishing real websites from fake ones over 90% of the time..
Keywords — Phishing, Detection, Phishing Websites, Legitimate, Machine Learning.

I. INTRODUCTION

Phishing is the most unsafe criminal activity in cyberspace. Now a days, most of the users go online to access the information and services provided by government and financial institutions, there has been a remarkable increase in phishing attacks for the past few years.Various methods are used by phishers to attack the vulnerable users such as messaging, VOIP, spoofed link,email and counterfeit websites. counterfeit websites are those which look like a genuine website in terms of layout and content. Even, the content of these counterfeit websites would be identical to their legitimate websites.

The United States Computer Emergency Readiness Team (US-CERT) defines “phishing as a form of social engineering that uses e-mails or malicious websites to solicit personal information from an individual or company by posing as a trustworthy organization or entity”. Even though organizations are educating their employees about how to recognize phishing emails or links to help and protect against the above types of attacks, software such as HTTrack is easily available for the users to duplicate entire websites for their own motives. As a result, even trained users can still be tricked into revealing private or sensitive information by interacting with a malicious website that they believe to be legitimate/trustworthy.

Phishing websites settle a variety of signals within its content-parts as well as the browser-based security indicators provided in the website. Several solutions have been proposed to tackle phishing. In Spite of everything, there is no single magic bullet that can solve this threat completely.

The discussed problem implies that computer-based solutions for guarding against phishing attacks are needed along with user education. Such a solution would enable a computer to have the ability to identify malicious/fake websites in order to prevent users from interacting with them. One general approach to recognizing malicious phishing websites relies on their Uniform Resource Locators (URLs). A URL is a global address of a document in the World Wide Web, and it serves as the primary means to locate a document or content on the Internet. Even in cases in which the content of a website is duplicated, the URLs could still distinguish real sites from fake ones.

One solution approach to detect phishing websites is to use a blacklist of malicious URLs developed by anti-virus groups. The problem with this approach is that the blacklist cannot be in-depth because new ill URLs keep growing continuously. Thus, approaches should be developed that can spontaneously classify a new or previously unseen URL as either a phishing site or a legitimate one. Such solutions are typically machine-learning based approaches where a system can categorize new phishing sites through a model developed using training sets of known attacks

II. LITERATURE REVIEW

In this section , the work done by others using different techniques to achieve the maximum accuracy result and improve the whole system will be discussed.

Fadi Thabtah [2] experimentally compared large numbers of Machine Learning techniques on real phishing datasets and with respect to different metrics. The main purpose of this comparison is to reveal the advantages and disadvantages of ML predictive models and to explain their actual performance in phishing attacks. The experimental results show that using approach models are more appropriate as anti phishing solutions as they are simple yet effective.

Muhemmet Baykara [3] proposed an application solution for phishing attacks which is known as "Anti Phishing Simulator", it gives information about the detection problem and how to detect phishing emails. Spam emails are added to the database by Bayesian algorithm. Phishing attackers use JavaScript to replace a legitimate URL of the URL onto the browser's address bar. The recommended approach in the study is to use the text of the email as a keyword only to perform complex word processing. "Anti Phishing Simulator" was developed to check the content and decide whether the message contained phishing elements.

Naghmeh Moradpoor [4] proposed a neural network-based model for detection and classification of phishing emails. It uses real kind emails from the "SpamAssassin" dataset and real phishing emails from "Phishcorpus" dataset. Python and MATLAB are used to measure the accuracy, true-positive rate, false positive-rate, network performance, and error histogram.

Mustafa Aydin [5] proposed a classification algorithm for phishing website detection by extracting website's URL features and analyzing subset based on feature selection methods. It implements feature extraction and selection methods for the detection of phishing websites. The extracted features about the page URLs and feature matrix are categorized into five different analyses as Alpha-numeric Character Analysis,Security Analysis, Keyword Analysis, Domain Identity Analysis and Rank Based Analysis. Most of these features are the textual properties of the URL itself and others based on third parties services.

Authors [6]]proposed a model with an answer for recognizing phishing sites by utilizing URL identification strategy using Random Forest algorithm. Their model has three stages, namely Parsing, Heuristic Classification of data, Performance Analysis.They used parsing to analyze feature sets.

Authors [7] proposed various algorithms and approaches to detect phishing websites by several researchers in Machine Learning. On reviewing the papers, they came to a conclusion that most of the work was done by using familiar machine learning algorithms like Naïve Bayesian, SVM, Decision Tree and Random Forest.

III. METHODOLOGY

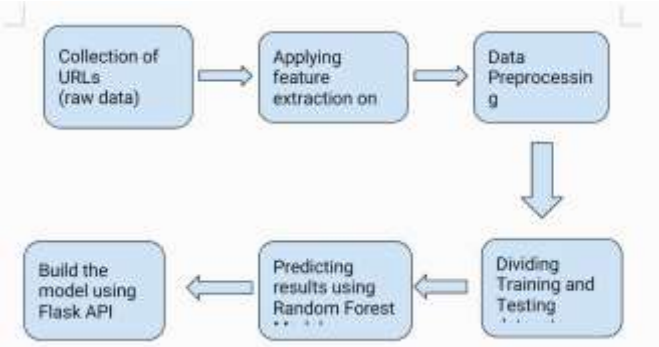

Figure 1: Block Diagram

Data Preprocessing

A raw dataset has been taken which consists of few website links where some of them are phishing websites and remaining are legitimate websites.

Extract the features from the data based on conditions and then a model can be built.

The data should get splitted based on the parts of the URL.

A typical URL could have the form "http://www.eample.com/index.html", where(http)indicates a protocol,(www.example.com)indicates a hostname,(index.html) indicates a file name.

Hence the final data after splitted based on the parts of the URL is:

	Protocol	domain_name	address
0	http	www.emuck.com:3000	archive/egan.html
1	http	danoday.com	summit.shtml
2	http	groups.yahoo.com	group/voice_actor_appreciation/links/events_an...
3	http	voice-international.com	
....	.....	.....	.....

Feature Extraction

Applying Feature Extraction on the splitted data(based on the parts of the url)

Feature -1:

Long URL to hide the suspicious Part.

Basically an URL consists of length,a maximum of 54 characters. If an URL is of length greater than 54 characters,it is termed as a long url which hides the suspicious part.

Hence the data can be divided into 2 categories based on the length, if it is above 54 characters then it is a phishing website or legitimate website.

Feature -2:

URL's having "@" Symbol.

Using "@" symbol in the URL leads the browser to ignore everything preceeding"@"symbol and real address often follows"@"symbol

If "@" symbol is in the URL,then it is termed as Phishing ,otherwise Legitimate.

Feature -3:

Redirecting using"//"

The existence of "//" within the URL path means that the user will be redirected to another website.

An example of such URL's is: "http://www.legitimate.com//http://www.phishing.com".

We examine the location where the "//" appears. So if the URL starts with "HTTP", that means the "//" should appear in the sixth position.

However, if the URL employs "HTTPS" then the "//" should appear in seventh position.Based on the last occurence of "//" ,it can be divided,whether it is phishing or legitimate.

Feature -4:

Adding Prefix or Suffix Separated by (-) to the Domain

The dash symbol(-)is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage.

For example http://www.Confirme-paypal.com/. Hence if an URL is having(-) symbol,it is categorized as a phishing otherwise legitimate website.

Feature-5:

Sub-Domain and Multi Sub-Domains

The legitimate URL link has two dots in the URL since we can ignore typing "www.".However, if the dots are greater than three it is classified as a "Phishing website" since it will have multiple sub-domains.

Feature -6:

Using the IP Address

Sometimes an IP address can be used as an alternative to the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information.

Sometimes,the IP address can be transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Hence based on the IP address,data can be divided.

Feature-7:

URL Shortening Services "TinyURL"

It is a method in the "World Wide Web" in which an URL may be made considerably smaller in length and still lead to the required webpage.

This is accomplished by means of an "HTTP Redirect" on a domain name that is short which links to the webpage that has a long URL.

For example,the URL "http://portal.hud.av.uk/" can be shortened to "bit.ly/19DXSk4".

Feature-8:

Existence of "HTTPS" Token in the Domain Part of the URL.

Phishers may add the "HTTPS" token to the domain part of a URL in order to trick users.For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

Such types of websites are termed as Phishing websites.

Feature-9:

Abnormal_URL

This feature can be extracted from WHOIS database.For a legitimate website, identity is typically part of its URL else it is Phishing.

Feature-10:

Google Index

This feature examines whether a website is in Google's index or not.

When a site has an index in Google,it can be displayed on the search results (Webmaster resources, 2014).

Usually,many phishing webpages may not be found on the Google index, hence they are merely accessible.

Feature-11:

Website Traffic

This is a feature, If a website has been visited by the user frequently,it has popularity based on the count of visitors. However,Phishing websites have a short life span,hence they would not have web traffic.

This feature can also separate the websites.

feature-12:

Domain Registration Length

Basically, based on the fact that a phishing website has a short period of time, it can be believed that trustworthy domains are regularly paid for several years in advance.

In our dataset, we find that the longest fraudulent domains have been used for one year only.

Feature-13:

This feature can be extracted from WHOIS database (Whois 2005).

Mostly phishing websites live for a short period of time. By reviewing the dataset,it can find that the minimum age of the legitimate domain is 5 months.

Feature-14:

DNS Record

If a website is not recognized by the WHOIS database (Whois 2005) or it has no records found for the hostname (Pan and Ding 2006).

If the DNS record is empty or not found then the website is classified as "Phishing", otherwise it is classified as "Legitimate".

Feature-15:

Statistical-Reports Based Feature

Several parties such as PhishTank and StopBadware formulated numerous statistical reports on phishing websites at every given period of time like monthly and quarterly.

After applying all the features on the raw dataset, minimized or free data sets will be generated. Then turn that dataset into a CSV file.

Machine Learning algorithms:

Apply machine learning algorithms like,Random forest and Decision tree classifier on the CSV file that got generated.

Firstly, data should get pre-processed.

Splitting the data into training and testing.Among both the cases,random forest gives the best accuracy,which could separate phishing and legitimate.

A flask has been developed for the application to give input as a link and get the output in the webpage whether it is phishing or legitimate ones

Table 1: Features that has been used to divide raw data set into a csv excel file

Long URL to hide the suspicious Part
URL's having "@" Symbol.
Redirecting using"//"
Adding Prefix or Suffix Separated by (-) to the Domain
Sub-Domain and Multi Sub-Domains
Using the IP Address
URL Shortening Services "TinyURL"
Existence of "HTTPS" Token in the Domain Part of the URL.
Abnormal_URL
Google Index

Website Traffic
Domain Registration Length
WHOIS database:
DNS Record
Statistical-Reports Based Feature

Legitimate URLs:

http://www.uvm.edu/~hearts/
http://www.uvm.edu/~boulder/
http://www.uvm.edu/~skiclub/
http://www.uvm.edu/greening/
http://www.uvm.edu/~uvmsga/
http://www.uvm.edu/~rlweb/ira/
http://www.emba.uvm.edu/~asem
http://www.emba.uvm.edu/~asce/
http://www.emba.uvm.edu/~asme/
http://www.emba.uvm.edu/~cssa/
http://www.uvm.edu/~geogclub/
http://www.uvm.edu/~goodrich/
http://pss.uvm.edu/hortclub/
http://www.uvm.edu/~phibeta/
http://universitycommunications.uvm.edu/fall99releases/webclimbthewalls.htm
http://www.uvmgreeklife.com/
http://www.agrinvt.org/
http://www.uvm.edu/~fiji2
http://www.uvm.edu/~sigep/
http://www.uvm.edu/~aphio/
http://www.uvm.edu/~sigmaphi/home.htm
http://www.uvm.edu/~uvmaz/
http://uvm.phideltatheta.org/
http://www.uvm.edu/~tridelt/
http://www.uvm.edu/~uvmivcf/
http://www.uvmhillel.org/
http://www.uvm.edu/~ohpr/
http://www.uvm.edu/~helix/
http://www.uvm.edu/~vlrs/
http://pss.uvm.edu/dept/hort_farm/
http://www.uvm.edu/health/
http://www.uvm.edu/~tpswww/
http://uds.uvm.edu/
http://www.uvm.edu/~access/
http://www.uvm.edu/~uvmiac/
http://www.uvm.edu/~career
http://siri.uvm.edu/
http://esf.uvm.edu/
http://www.uvm.edu/~wrtngctr/
http://www.vermontlaw.edu/
http://www.vjel.org
http://www.woodbury-college.edu/
http://www.gwvirginia.gwu.edu/
http://www.cdu.edu/
http://www.umtweb.edu/
http://www.virginia529.com/
http://www.uacp.org
http://www.cordobauniversity.org/
http://www.artinstitutes.edu/arlington/
http://www.apprenticeschool.com
http://www.gobuilders.com/
http://www.abbc.edu/
http://www.averett.edu/

Phishing URL's

http://asesoresvelfit.com/media/datacredito.co/
http://hissoulreason.com/js/homepage/home/
http://unauthorizd.newebpage.com/webapps/66fbf/
http://133.130.103.10/23/
http://dj00.co.vu/css/?bsoul=Qg@xIHW%//yh/en/?i=34453&i=34453
http://133.130.103.10/21/logar/
http://httpssicredi.esy.es/servico/sicredi/validarclientes/mobi/index.php
http://gamesaty.ga/wp-content///yh/en/?i=31416&i=31416
http://luxuryupgradepro.com/ymailNew/ymailNew/
http://133.130.103.10/1/
http://133.130.103.10/24/sicredi/psmlId/31/paneid/index.htm
http://smscaixaacesso.hol.es
http://133.130.103.10/7/SIIBC/siwinCtrl.php
http://tinyurl.com/kjmmw57
http://wrightlandscapes.org/no/T/Y1.html
http://ginatringali.com//al/alibaba21012015/alibaba21012015/666/index.html
https://staticmail.000webhostapp.com/
http://umeda.com.br/bba/BOA/home/
http://krishworldwide.com/BackUp/under/js/ayo1/index.html
http://yahoo.co.in/email_open_log_pic.php?mid=9f8fd3e2a108a256bff453d09c965c25&s=a
http://www.avcc.ac.in/fonts/1/wropboxp/login.htm

IV. RESULTS AND CONCLUSIONS

This paper focuses on detecting phishing website URLs by considering the following features Protocol, path, domain, having_ip ,len_url ,having_at_symbol ,redirection_symbol , prefix_suffix_separation , sub_domains , tiny_url , Web_traffic , domain_registration_length , dns_record , statistical_report , age_domain , http_tokens . Feature Extraction is done on the urls and the extracted data set is classified into phishing or legitimate using the Random Forest classification model .

This paper aims to improve the methods or enhance the method to predict whether the given website is a phishing website or legitimate website using machine learning technology. It achieved 90.14% detection accuracy using the Random Forest algorithm with the lowest false positive rate.

Also the result shows that classifiers give better performance if we used more data as the training data

Figure 2: Home Page

Figure 3: A URL is given as input and it is detected to be legitimate

Figure 4: A URL is given and it is detected to be Phishing

V. FUTURE WORK

The research work presented here has some limitations and it can be extended further. The limitation is that we considered a small data set that contains 2017 URLs, and there are 15 features for each URL. The results motivate future works to add more features to the dataset, which could improve the performance of these models, hence it could combine machine learning models with other phishing detection techniques to obtain better performance. Besides, we will explore in order to propose and develop a new mechanism to extract new features from the website to keep up with new techniques in phishing attacks and also it is important to detect phishing websites in real time whenever the user tries to enter rather than the user checking the URL with our application.

REFERENCES

[1] Anjum N. Shaikh, Antesar M. Shabut and M. A. Hossain, " A literature review on Phishing Crime, Prevention Review and Investigation of gaps", 2016 10th International Conference on Software, Knowledge, Information Management & Applications.

[2] Neda Abdelhamid, Fadi Thabtah and Hussein Abdel-jaber, "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features", IEEE Int. Conf. on Intelligence and Security Informatics (ISI), pages 72–77, 2017.

[3] Muhammet Baykara, Zahit Ziya Gürel, Detection of phishing attacks,2018.

[4] Naghmeh Moradpoor, Employing Machine Learning Techniques for Detection and Classification of Phishing Emails, July 2017.

[5] Mustafa Aydin, Nazife Baykal, Feature Extraction and Classification Phishing Websites Based on URL, 2015.

[6] 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.

[7] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning" in 2019 at International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11.

.