Fake Website Detector

Kamlesh Gupta[1], Rohit Gharal[2], Gunasekar Sivasankar[3], Pratap Nair[4]
[1,2,3]*Excelssior Education Society's, KC College of Engineering and Management Studies and Research, Kopri, Thane (East), Mumbai, Maharashtra, India*
[4]*Assistant Professor, KC College of Engineering and Management Studies and Research, Kopri, Thane (East), Mumbai, Maharashtra, India*

***Abstract*: Web service is a communication protocol and software between two electronic devices over the Internet. It extends the World Wide Web infrastructure to provide the methods for an electronic device to connect to other electronic devices. With the rapid development of the Internet and the increasing popularity of electronic payment in web service, Internet fraud and web security have gradually been the main concern of the public.**

Phishing is an illegal activity wherein people are misled into the fake sites by using various fraudulent methods. Web Phishing is a way of such activity, which uses social engineering technique through short messages, emails, and other social media to induce users to visit fake websites to get sensitive information like their private account, token for payment, credit card information, and so on. Hence it leads to information disclosure and property damage. As the technology advances, it needs a better model to secure people from such cyber attack.

***Keywords*: phishing, fake website, random forests classification, machine-learning.**

I. INTRODUCTION

As the time is running, new technology is being developed. We always face a new technology in our day to day life. As the technologies is being advanced, every time an unknown threat challenge to human being. Although these technologies are important to human life, they also create many problems in their life. Among these problems one is phishing.

Few years back when there was no internet or it was in developing phase, the hacker uses different method other than phishing such as social engineering, vishing (voice over phishing), mishing (mobile phishing) etc. Since there wasn't much access or exposure to online procedures, online dealings or transactions and lack of internet access, there was very minimal threat to the then already existing online systems.

Since the last decade, there has been a tremendous growth in information technology which has brought the most of the daily human routine come online; right from shopping to bank transactions. With the rapid development of the Internet and the increasing popularity of electronic payment in web service, internet fraud and web security have gradually been the main concern of the public. Phishing is such form of fraud in which the attacker tries to get sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other way of communication channels. Web Phishing is an illegal activity in which hacker uses social engineering technique through observation, short messages, emails, social media to induce users to visit fake websites to get sensitive information like their private account, token for payment, credit card information, and so on. Phishing not only steal sensitive information but also it leads to defame one's life, information disclosure and property damage etc. Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.

The term 'Phishing' was coined by hackers who were stealing 'America On-Line' accounts by scamming passwords from unsuspecting its users in the year 1996. The ideology behind word 'Phishing' is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant social media site, which will take the user to hostile phishing websites.

There are various modes by which phishing can be carried out and hence there are different types of phishing like Email phishing, Vishing (voice over phishing), Smishing (Phishing via SMS), Whaling, Mishing (mobile phishing), social engineering, Spear phishing, Angler phishing etc. Usually, there are following phases in a typical phishing attack like social engineering, penetration (bait), observation (hook), mass broadcast, mature and account hijack. Sometimes they just make minor changes in URL of any reputed company and misled the user. See the below example:

https://WWW.facebook.com =>
https://WWW.facebok.com

This link redirect the user to malicious page which intimate user to fill their private data. Due to unnoticeable change in link, people generally ignore such things. It becomes more prominent when the link shown redirects the user to any payment page. A very good is of fraud in PayPal site.

As the time elapsing and new technologies coming every day, phishing attacks and their types grew in number and intensity too. Phishing attacks now target users of online banking, payment services such as PayPal, online e-commerce

sites and social media platform. With the prevalence of network, phishing has become one of the most serious security threats in modern society, thus making detecting and defending against web phishing an urgent and essential research task. Web phishing detection has been very crucial for both private users and enterprises.

In most of the phishing attacks, whether it is carried out by emails or any other medium, the main objective is to make the victim to follow a link that appears to go to a legitimate web resource but actually redirects the victim to a malicious web page. This suspicious web page looks very similar to original web page. Users are asked to fill their credential and once they fill and submit it, their private information is hacked. The simplest approach to tackle such phishing attack is to create a system which should be able to identify and warn user of such web page.

In this paper, we are focusing on detection of phishing websites by combination of URL based and content based detection. Some possible solutions to combat phishing were created, including specific legislation and technologies. In the earlier times, methods such as k-nearest neighbors, list-based approach, fuzzy logic and other mining and classification approaches were used for detection but as the intensity of the attack grew over the times more sophisticated algorithms and techniques were introduced to detect and prevent the attack. From a technical point of view, the detection of phishing generally includes the following categories: detection based on a black list and white list, detection based on Uniform Resource Locator (URL) features, detection based on web content, and detection based on machine learning. The anti phishing way using blacklist may be an easy way, but it cannot find new phishing websites. The detection on URL is to analyze the features of URL. The URL of phishing websites may be very similar to real websites to the human eye, but they are different in IP. The content-based detection usually refers to the detection of phishing sites through the elements of pages such as form information, field names, and resource reference etc.

II. METHODOLOGY

Our goal is to classify a web page as legitimate or fake; in this section, we first describe the datasets we used for our training and testing, then we present the model selection and evaluation process and lastly the proposed architecture approach by which we get desired output.

A. Dataset

We have taken database from 'Kaggle' named 'Phishing website dataset'. The resulting dataset consists of 11055 rows and 32 columns. After removing index and result column, the dataset have 30 optimized features of phishing website. These features are very essential for predicting the web pages. These features comprised of various parameters for URL based and content based detection. Features for URL based phishing detection are IP address, URL length, URL shortening service, at (@) symbol, double slash (//), Prefix_Suffix, sub domain, SSL certificate, domain registration length, favicon, port number and HTTPS token. Features for content based phishing detection are request URL, anchor tag URL, link in tag, SFH, submit to email, abnormal URL, redirect link, link pointing to other page, on mouseover, right click, popup windows and iframe. There are some features related to network such as age of domain, DNS record, web traffic, page rank, google index and statistical report. As the dataset is already clean, there is no need to perform cleanup process. Now all is good to proceed further.

B. Model Selection And Evaluation

We have started with training our dataset. We have trained our dataset by two ways. In first trial we have divided dataset into training dataset and testing dataset. In second trial we have selected complete dataset for training purpose and part of that for testing purpose. We also performed cross validation as this significantly reduces bias as we are using most of the data for fitting, and also significantly reduce variance as most of the data is also being used in validation set. This clear us which method can provide better accuracy.

A model is the heart of any artificial system. Therefore creating an absolute model is very necessary. As of now there are many classification algorithms available. Few of them are naïve bayes, logistic regression, decision tree, support vector classifier, K-nearest neighbor, random forest etc. To create model, we have trained and tested our dataset with these algorithm. Once we have computed a prediction, we found random forest classifier giving the highest accuracy. Below is the comparison table which gives a clear visualization of their accuracy.

Algorithm	Accuracy in %
Naïve Bayes	60
Logistic Regression	92
Support Vector Classifier	95
K-nearest Neighbor	97
Random Forest Classifier	98

After getting desired accuracy from random forest classifier, we created the model based on RF classifier. After this we have analyzed the model by both the approach of dividing dataset. It can be noted that accuracy of model doesn't vary significantly by changing dataset training approach.

C. Proposed Architecture

The core of our system is the model used as backed service. The goal of our system is to alert it's user about visiting web page and securing their credential. The system is

comprised of layout (frontend), model (backend) and dataset. Each part plays an essential role in making the system work.

As stated earlier, we have combined URL-based detection approach and content-based detection approach. Starting with the input, we have URL as well as content of web page as an input. We also gathered some information regarding network of webpage such as IP address, port number, SSL certificate, HTTPS token, age of domain, DNS record, web traffic, page rank, google index and statistical report. Using JavaScript method implemented in extension we get URL and content of web page and fetch it to backend using methods available in Ajax. In backend service we receive these input using methods available in PHP.

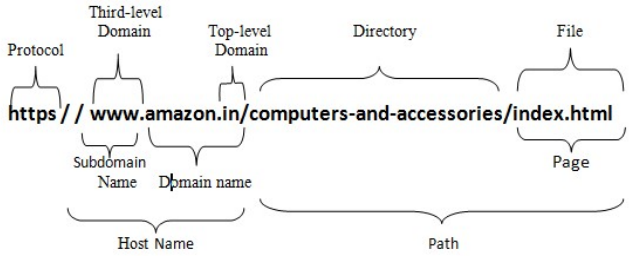

Figure 1: Features in URL

Once we got URL, we extracted its parameter according to features available to us. To retrieve the content of a webpage, we applied some simple heuristics: we removed the CSS and JavaScript content from the page, then we extracted the texts, links, forms value, iframe etc contained in the remaining HTML tags. Finally, we performed the post classification using a random forest classifier model. To process the data, we fetch the data to model. After successfully processing we get the desired output. Now this result is fetched to extension to alert user as output.

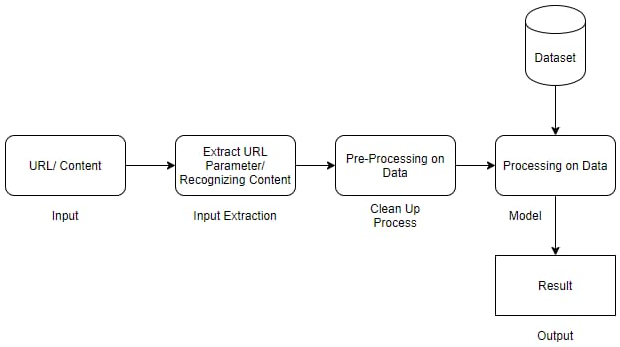

Figure 2: Generalized view of system

III. CONCLUSION AND FUTURE WORK

The system in the form of extension has been proposed that can detects phishing websites by using random forests as classification algorithm with scripting language. Here we have used dataset taken from 'Kaggle' that has 30 features. Out of those 30 features there are group of features that are most suitable for detection of phishing websites. Upon varying these features, the accuracy of model also varies. The performance metrics along with our literature survey and algorithm accuracy comparison also proved the accuracy level of random forest to be the highest around 98% and thus Random Forests algorithm were chosen for classification.

The proposed system as described reads URL from browser and content of that page as input and feed them to model. The model then process on input and fetches the promising result to alert user.

There has been lots of research done in this area that shows there is no single technique, that is enough to detect all types of phishing attack till now and with the upcoming technology, the type and number of phishing attacks are expected to increase. For these, the browsers have to be made capable enough to setup methods that detect and warn of potential phishing attacks. Future work will aim to develop a system that is extension that can learn by itself about new types of phishing attacks by adding a more enhanced feature to the detection process. It will also enable user to take action by them accordingly.

ACKNOWLEFGEMENT

The authors would like to thank prof. Pratap Nair for the precious support and the time dedicated to the review of the work.

REFERENCES

[1] Ludl, C., McAllister, S., Kirda, E., & Kruegel, C. (2007). On the effectiveness of techniques to detect phishing sites. In Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 20-39). Springer Berlin Heidelberg.

[2] Anti-Phishing Working Group Phishing, (2014). Anti Phishing Working Group Phishing Trends Report. [Online] Available at: https://apwg.org/ [Accessed 30 Mar. 2015].

[3] Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. Communications of the ACM, 50(10), 94-100.

[4] Zhang, Y., Hong, J. I., & Cranor, L. F. (2007, May). Cantina: a content-based approach to detecting phishing web sites. In Proceedings of the 16th international conference on World Wide Web (pp. 639-648). ACM.

[5] Dunlop, M., Groat, S., & Shelly, D. (2010, May). Goldphish: Using images for content-based phishing analysis. In Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on (pp. 123-128). IEEE.

[6] The center frequency of the previous filter and the upper boundary is situated at the center frequency of the next filter.

[7] The maximum response of a filter, that is, the top vertex of the triangular filter, is located at the filter's center frequency and is normalized to unity.

AUTHOR BIOGRAPHIES

Kamlesh Gupta

Born in Bhadohi, Uttar Pradesh, India on 10/03/1996.

The author is currently pursuing Bachelors of Engineering in the stream of Computer Science from Excelssior Education Society's KC College of Engineering and Management Studies and Research and will earn his UG degree by 2020.

Rohit Gharal

Born in Satara, Maharashtra, India on 30/07/1997.

The author is currently pursuing Bachelors of Engineering in the stream of Computer Science from Excelssior Education Society's KC College of Engineering and Management Studies and Research and will earn his UG degree by 2020.

Gunasekar Sivasankar

Born in Gudiyatham, Tamilnadu, India on 11/10/1998.

The author is currently pursuing Bachelors of Engineering in the stream of Computer Science from Excelssior Education Society's KC College of Engineering and Management Studies and Research and will earn his UG degree by 2020.