International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Detection of Phishing Websites using Machine Learning

Atharva Deshpande
4th year, BE Computer Engineering Student,
Vidyavardhini's College of Engineering & Technology
Vasai, Mumbai

Omkar Pedamkar
4th year, BE Computer Engineering Student,
Vidyavardhini's College of Engineering & Technology
Vasai, Mumbai

Nachiket Chaudhary
4th year, BE Computer Engineering Student,
Vidyavardhini's College of Engineering & Technology
Vasai, Mumbai

Dr. Swapna Borde
Dept. of Computer Engineering,
Vidyavardhini's College of Engineering & Technology
Vasai, Mumbai

***Abstract*— Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishers use the websites which are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning.**

Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents.

Here, we explain phishing domain (or Fraudulent Domain) characteristics, the features that distinguish them from legitimate domains, why it is important to detect these domains, and how they can be detected using machine learning and natural language processing techniques.

Keywords*— *Phishing, Personal information, Machine Learning, Malicious links, Phishing domain characteristics.

I. INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. Phishing may be a style of broad extortion that happens once a pernicious web site act sort of a real one memory that the last word objective to accumulate unstable info, as an example, passwords, account focal points, or MasterCard numbers. all the same, the means that there square measure some of contrary to phishing programming and techniques for recognizing potential phishing tries in messages and characteristic phishing substance on locales, phishes think about new and crossbreed procedures to bypass the open programming and frameworks. Phishing may be a fraud framework that uses a mixture of social designing what is additional, advancement to sensitive and personal data, as an example, passwords associate degree open-end credit unpretentious elements by presumptuous the highlights of a reliable individual or business in electronic correspondence. Phishing makes use of parody messages that square measure created to seem substantial and instructed to start out from true blue sources like money connected institutions, online business goals, etc, to draw in customers to go to phony destinations through joins gave within the phishing websites.

II. STATE OF THE ART (LITERATURE SURVEY)

H. Huang et al., (2009) proposed the frameworks that distinguish the phishing utilizing page section similitude that breaks down universal resource locator tokens to create forecast preciseness phishing pages normally keep its CSS vogue like their objective pages.

S. Marchal et al., (2017) proposed this technique to differentiate Phishing website depends on the examination of authentic site server log knowledge. An application Off-the-Hook application or identification of phishing website. Free, displays a couple of outstanding properties together with high preciseness, whole autonomy, and nice language-freedom, speed of selection, flexibility to dynamic phish and flexibility to advancement in phishing ways.

Mustafa Aydin et al. proposed a classification algorithm for phishing website detection by extracting websites' URL features and analyzing subset based feature selection methods. It implements feature extraction and selection methods for the detection of phishing websites. The extracted features about the URL of the pages and composed feature matrix are categorized into five different analyses as Alpha-numeric Character Analysis, Keyword Analysis, Security 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.

Samuel Marchal et al. presents PhishStorm, an automated phishing detection system that can analyze in real time any URL in order to identify potential phishing sites. Phish storm is proposed as an automated real-time URL phishingness rating system to protect users against phishing content. PhishStorm provides phishingness score for URL and can act as a Website reputation rating system.

Fadi Thabtah et al. experimentally compared large numbers of ML techniques on real phishing datasets and with respect to different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more appropriate as anti-phishing solutions. Muhemmet Baykara et al. proposed an application which is known as Anti Phishing Simulator, it gives information about the detection problem of phishing and how to detect phishing emails. Spam emails are added to the database by Bayesian algorithm. Phishing attackers use JavaScript to place a legitimate URL of the URL onto the browsers address bar. The recommended approach in the study is to use the text of the e-mail as a keyword only to perform complex word processing.

III. PROJECT DESCRIPTION

We have developed our project using a website as a platform for all the users. This is an interactive and responsive website that will be used to detect whether a website is legitimate or phishing. This website is made using different web designing languages which include HTML, CSS, Javascript and Django.

The basic structure of the website is made with the help of HTML. CSS is used to add effects to the website and make it more attractive and user-friendly. It must be noted that the website is created for all users, hence it must be easy to operate with and no user should face any difficulty while making its use. Every naïve person must be able to use this website and avail maximum benefits from it.

The website shows information regarding the services provided by us. It also contains information regarding ill-practices occurring in today's technological world. The website is created with an opinion such that people are not only able to distinguish between legitimate and fraudulent website, but also become aware of the mal-practices occurring in current world. They can stay away from the people trying to exploit one's personal information, like email address, password, debit card numbers, credit card details, CVV, bank account numbers, and the list goes on.

The dataset consists of different features that are to be taken into consideration while determining a website URL as legitimate or phishing.

The components for detection and classification of phishing websites are as follows:

1. Address Bar based Features
2. Abnormal Based Features
3. HTML and JavaScript Based Features
4. Domain Based Features

1. Address Bar based Features

1. Using the IP address

If IP address is used instead of domain name in the URL e.g. 125.98.3.123 the user can almost be sure someone is trying to steal his personal information.

2. Long URL to hide the Suspicious Part

Phishers can use long URL to hide the doubtful part in the address bar.

3. Using URL shortening services TinyURL

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

4. URLs having @ symbol

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

5. Redirecting using //

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

6. 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.

7. Sub Domain and Multi Sub Domains

Let us assume we have the following link: http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (ccTLD).

8. HTTPs (Hyper Text Transfer Protocol with Secure Sockets Layer)

The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough.

9. Domain Registration Length

Based on the fact that a phishing website lives for a short period of time, we believe 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.

10. Favicon

A favicon is a graphic image (icon) associated with a specific webpage.

11. Using Non-Standard Port

This feature is useful in validating if a particular service is up or down on a specific server.

12. The existence of HTTPS Token in the Domain Part of the URL

The phishers may add the HTTPS token to the domain part of a URL in order to trick users.

2. Abnormal Based Features

1. Request URL

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain.

2. URL of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as Request URL.

3. Links in <meta>, <Script> and <Link> tags

Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script; and <Link> tags to retrieve other web resources.

It is expected that these tags are linked to the same domain of the webpage.

4. Server From Handler(SFH)

SFHs that contain an empty string or about:blank are considered doubtful because an action should be taken upon the submitted information.

5. Submitting Information to Email

Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the users information to his personal email.

6. Abnormal URL

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

3. HTML and JavaScript Based Features

1. Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. Status Bar Customization

2. Disabling Right Click

Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. This feature is treated exactly as Using onMouseOver to hide the Link.

3. Using Pop-Up Window

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window.

4. IFrame Redirection

IFrame is an HTML tag used to display an additional webpage into one that is currently shown.

4. Domain Based Features

1. Age of Domain

This feature can be extracted from WHOIS database. Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6 months.

2. DNS Record

For phishing websites, either the claimed identity is not recognized by the WHOIS database or no records founded for the hostname. If the DNS record is empty or not found then the website is classified as Phishing, otherwise it is classified as Legitimate.

3. Website Traffic

This feature measuresthe popularity of the website by determining the number of visitors and the number of pages they visit.

4. Page Rank

PageRank is a value ranging from 0 to 1. PageRank aims to measure how important a webpage is on the Internet.

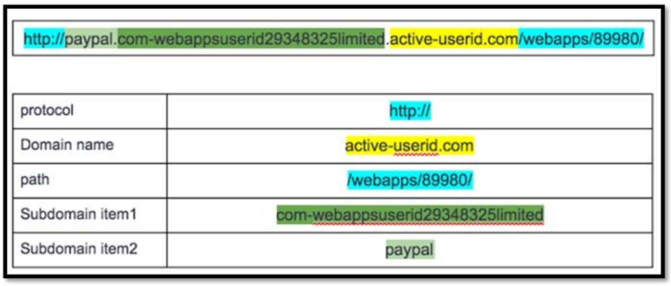
5. Google Index

This feature examines whether a website is in Googles index or not. When a site is indexed by Google, it is displayed on search results.

6. Number of Links Pointing to Page

The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain.

7. Statistical-Reports Based Feature



http://paypal.com-webappsuserid29348325limited.active-userid.com/webapps/89980/

protocol	http://
Domain name	active-userid.com
path	/webapps/89980/
Subdomain item1	com-webappsuserid29348325limited
Subdomain item2	paypal

Fig.1. URL parts and features

IV. ALGORITHMS USED

Two algorithms have been implemented to check whether a URL is legitimate or fraudulent.

Random forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy. Creation of trees is based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for classification.

Decision tree begins its work by choosing best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label.

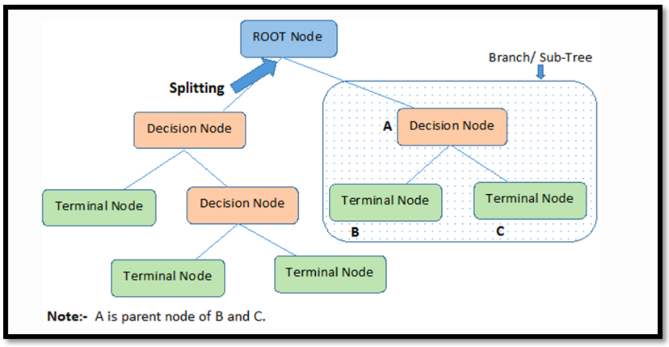


Fig.2. Decision Tree Algorithm working

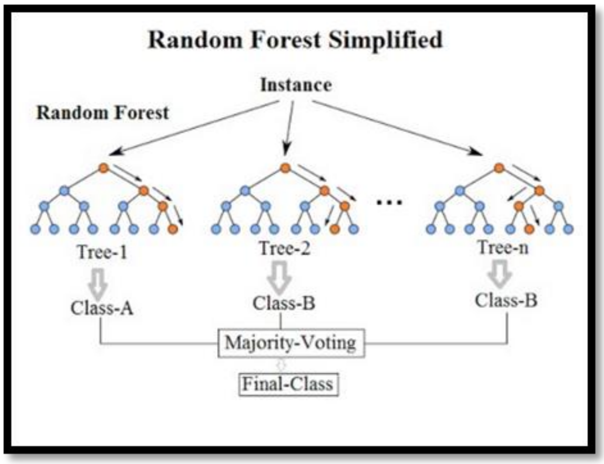


Fig.3. Random Forest Algorithm working

V. PROJECT REQUIREMENTS

Hardware Requirements:-

- 2GB RAM (minimum)
- 100GB HDD (minimum)
- Intel 1.66 GHz Processor Pentium 4 (minimum)
- Internet Connectivity

Software Requirements:-

- WINDOWS 7 or higher
- Python 3.6.0 or higher
- Visual Studio Code
- Django
- HTML
- Dataset of Phishing Websites

VI. WORKING

- We have collected unstructured data of URLs from Phishtank website, Kaggle website and Alexa website, etc.
- In pre-processing, feature generation is done where nine features are generated from unstructured data. These features are length of an URL, URL has HTTP, URL has suspicious character, prefix/suffix, number of dots, number of slashes, URL has phishing term, length of subdomain, URL contains IP address.
- After this, an organized dataset is made in which each detail incorporates the paired (0,1) which is then passed to the various classifiers.
- Next, we train the three unique classifiers and analyse their presentation based on exactness two classifiers utilized are Decision Tree and Random Forest algorithm.
- At that point, the classifier identifies the given URL dependent on the preparation information that is if the site is phishing it prompts the user that the website is phished and if genuine, it prompts the user that the website is legitimate.
- We look at the exactness of various classifiers and discovered Random Forest as the best classifiers which gives the most extreme precision.

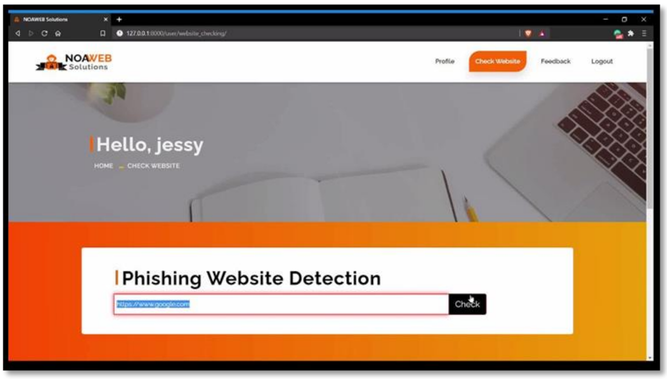
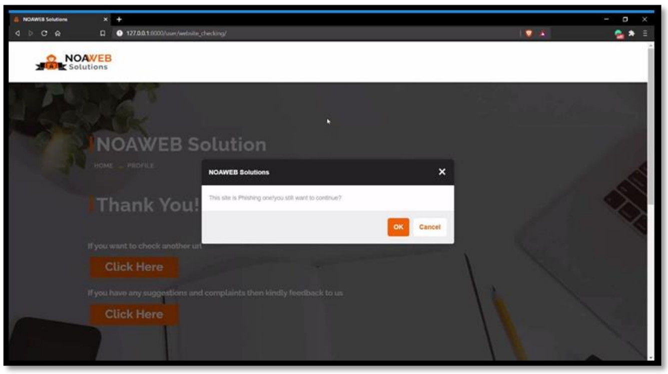


Fig.4. UI of Website Checking Portal

However, if the URL entered by a user is found to be a phishing website, a small pop-up will appear on the screen to warn the user regarding this malicious website. There are times when a user needs to access some data on that website, so he/she can select a 'CONFIRM' option to open the website, otherwise he/she will be sent back to the above webpage.



This site is Phishing one!you still want to continue?

If you want to check another url

If you have any suggestions and complaints then kindly feedback to us

Fig.5. Alert Warning for Fraudulent Websites

VII. RESULTS

Scikit-learn tool has been used to import Machine learning algorithms. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers.

Performance of classifiers has been evaluated by calculating classifier's accuracy score.

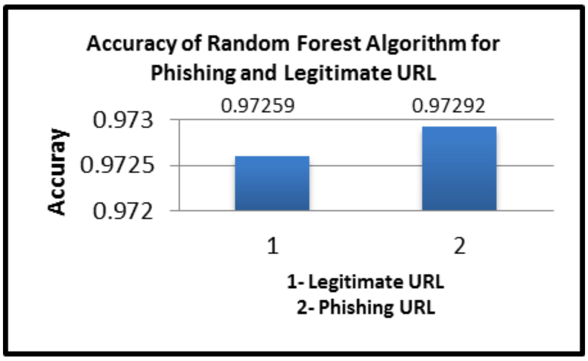


Fig.6. Accuracy with Random Forest Algorithm

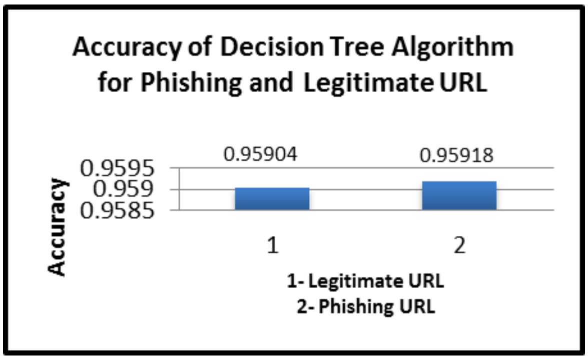


Fig.7. Accuracy with Decision Tree Algorithm

VIII. CONCLUSION

Thus to summarize, we have seen how phishing is a huge threat to the security and safety of the web and how phishing detection is an important problem domain. We have reviewed some of the traditional approaches to phishing detection; namely blacklist and heuristic evaluation methods, and their drawbacks. We have tested two machine learning algorithms on the 'Phishing Websites Dataset' and reviewed their results. We then selected the best algorithm based on it's performance and built a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. We have detected phishing websites using Random Forest algorithm with and accuracy of 97.31%. For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

IX. FUTURE SCOPE

Although the use of URL lexical features alone has been shown to result in high accuracy (~97%), phishers have learned how to make predicting a URL destination difficult by carefully manipulating the URL to evade detection. Therefore, combining these features with others, such as host, is the most effective approach .

For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

X. REFERENCES

[1] J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.

[2] Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[3] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[4] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[5] 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.

[6] K. Shima et al., "Classification of URL bitstreams using bag of bytes," in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[7] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature Selection for the Prediction of Phishing Websites," 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[8] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis," 2017.

[9] L. MacHado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.

[10] A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine leaning," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.