PHISHING WEBSITE DETECTION

Akanksha S Gatty, Poorvi M Rao, R Devika, Sanjana Arun Prabhas,

Information Science and Engineering Department, AJ Institute of Engineering and Technology

Abstract- Phishing is the technique of extracting user credentials and sensitive data from users by masquerading as a genuine website. In phishing, the user is provided with a mirror website that is identical to the legitimate one but with malicious code to extract and send user credentials to phishers. Blacklist, heuristic, visual scan and machine learning are a few anti-phishing strategies. Our proposed system does the task of extracting the various features from the URL, to check whether the site is a fraud site or not. Machine learning is a reliable method to recognize phishing.

I. INTRODUCTION

Nowadays, a large number of individuals are aware of the benefits of using the internet for a variety of tasks, including online banking, online bit payments, online cell recharging, and online shopping. The widespread usage of technology exposes customers to several security risks, including criminality. Spam, fraud, cyber terrorism, and phishing are only a few examples of the numerous cybercrimes that are frequently committed. Phishing is a brand-new cybercrime that is currently quite popular. Phishing may be defined as a deceptive method of obtaining or retrieving private information by deceiving a victim into thinking that the con artist is a trustworthy individual in whom the victim can place their faith. A phishing assault is frequently classified as an associate engineer attack or a social attack. Phishing assaults target people on a daily or frequent basis. Even some of the biggest organizations in the world are not immune to the increasingly sophisticated phishing attempts, which now average well over 1,000 per month. Phishers create websites that mimic legitimate websites and send emails that impersonate users in order to steal their private information, such as usernames, passwords, and financial information, for a variety of purposes. Phishing can be defined as a deceptive method of obtaining or retrieving sensitive personal information by deceiving an unnamed person into thinking the con artist is a good person that the unnamed person can rely on. Scammers target thoughtless and complacent people because they are simple prey for predators and are easy prey for them. Phishing attacks are frequently also referred to as social engineering attacks or social attacks. Phishing assaults target people on a daily or frequent basis. In recent years, phishing attempts have grown and multiplied greatly, becoming more sophisticated and cynical in nature. Even some of the biggest organizations in the world are not immune to the increasingly sophisticated phishing attempts, which now average well over 1,000 per month. A recent survey found that more than 65% of organizations would have to deal with phishing attacks in 2020, 30% of targeted users had opened phishing messages, 32% of data threats were discovered in 2019, phishing was the primary cause of 78% of cyber-espionage incidents, and 51% of phishing attacks contained links to malware. According to an IBM study, the price of a data breach caused by a phishing assault might reach $4 million; however, the reported sum is insufficient to fully account for the effects and monitory losses caused by phishing attacks. According to one of the FBI's reports on internet crime, US businesses alone had to pay out more than $1.2 billion as a result of business e-mail compromise attacks; scammers using fake gift cards, one type of spear phishing attack where the gift card is supposedly being sent, has been costing more than $70 million a year; another type of phishing attack is direct deposit phishing where the scammers are able to retrieve other people's employee portal information by stealing their salaries which accounts for more than $100 million loss to businesses.

II. METHODOLOGY

A. Features

1. *Subdomains:*

   If there are more than three dots in the URL using this technique, it is phishing.

2. *Domain:*

   The domain name or network address of the URL is returned by this method, which also accepts a URL string as its input.

3. *IP Address in the URL:*

   It looks at the URL to determine whether an IP address is there. In URLs, IP addresses may take the place of domain names. We may be assured that someone is trying to steal personal information with this URL if the domain name is replaced with an IP address. When an IP address appears in the domain portion of a URL, the value for this characteristic is either 1 (phishing) or 0 (legal).

4. *"@" Symbol in URL:*

   It verifies whether the URL contains the '@' symbol. Using the "@" sign in a URL causes the browser to ignore everything before the "@," and the actual address frequently comes after the "@." The value assigned to this feature is 1 (phishing) or 0 (legal) depending on whether the URL contains the symbol "@".

5. *Length of URL:*

   It determines the URL's length. Long URLs can be used by scammers to conceal the dubious portion in the address bar. In this project, a URL is classed as phishing if it has more than 54 characters but is otherwise legal. The value assigned

to this feature is 1 (phishing) or 0 (legal) depending on whether the length of the URL exceeds 54.

6. *Redirection "//" in URL:*

It examines the URL to see whether "//" is present. The visitor navigates to another website if the URL path contains the character "//". The "//" in a URL's location is computed. We discover that the "//" should be in the sixth position if the URL begins with "HTTP". However, if "HTTPS" is included in the URL, the "//" should be in the seventh place. The value assigned to this characteristic is either 1 (phishing) or 0 (legal) if the "//" character appears anywhere in the URL other than immediately following the protocol.

7. *"http/https" in Domain name:*

Verifies whether the domain portion of the URL contains the characters "http/https". In order to deceive users, phishers may append the "HTTPS" token to the domain portion of a URL. The value assigned to this feature is 1 (phishing) or 0 (legal) depending on whether the URL has "http/https" in the domain part.

8. *Using URL Shortening Services "TinyURL":*

The "World Wide Web" has a technology known as URL shortening that allows a URL to be significantly reduced in size while still directing to the desired web page. This is done by using a "HTTP Redirect" on a short domain name that links to the full URL of the web page. The value assigned to this feature is either 1 (phishing) or 0 (legal) depending on whether the URL uses shortening services.

9. *Prefix or Suffix "-" in Domain:*

It checks if there is a '-' in the URL's domain. Legitimate URLs rarely employ the dash symbol. Prefixes or suffixes, separated by (-), are often added to the domain name by phishers to give users the idea that they are visiting a reliable website. If the domain portion of the URL contains the symbol "-," the value assigned to this characteristic will either be 1 (phishing) or 0 (legal).

10. *DNS Record:*

Phishing websites either have no records for the host name or the claimed identity is not detected in the WHOIS database. The value assigned to this feature is either 1 (phishing) or 0 (legal) if the DNS record is null or cannot be located.

11. *Web Traffic:*

This function measures the popularity of the website by counting the number of visitors and the number of pages they view. However, because phishing websites only exist for a brief time, the Alexa database might not be able to identify them (Alexa the Web Information Company, 1996). Reviewing our data set, we see that, even in the most extreme cases, trustworthy websites were among the top 100,000. Additionally, it is considered "Phishing" if the domain receives no traffic or is not listed in the Alexa database. The value of this feature is 1 (phishing) if the domain rank is less than 100000; otherwise, it is 0 (legal).

12. *Age of Domain:*

To extract this attribute, use the WHOIS database. The majority of phishing websites are only active for a little time. For this initiative, a legal domain must have a minimum age of 12 months. Age in this context simply refers to the interval between creation and expiration times. The value of this attribute is 1 (phishing) if the domain's age exceeds 12 months, else 0 (legal).

13. *End Period of Domain:*

To extract this attribute, use the WHOIS database. The remaining domain time for this feature is determined by comparing the current time to the expiration time. For this project, the end time taken into account for the legal realm is six months or fewer. If the domain's end term is more than six months, the value of this attribute is 1 (phishing), otherwise it is 0 (legal).

14. *Abnormal URL:*

The function checks if the domain name is present in the URL. If it is, the function returns 0 indicating that the URL is legitimate. If the domain name is not present in the URL, the function returns 1 indicating that the URL is potentially a phishing website.

15. *Statistical Report:*

Using regular expressions, the function first extracts the host name from the input URL. The host name is then compared using regular expressions to see if it matches any of the established patterns. The function returns 1, indicating that the URL is not secure, if it does.

B. Data set used:

We gathered 6,34,626 genuine and phishing URLs; 999 of them contain phishing URLs, 101 have authentic art, 1017 contain legitimate URLs, and 632509 contain balanced URLs. Legitimate and authentic URLs make up balanced URLs.

III. RESULTS

The script reads two CSV files containing legitimate and phishing URLs into separate data frames using the Pandas library's pd.read_csv() function.The two data frames are joined into one data frame using the Pandas append() method. Some columns are discarded using the drop() method to prepare the data for training.The data are separated into training and test sets using the train_test_split() function of the scikit-learn module.CHANGE LEARN A random forest classifier is created using the RandomForestClassifier() method. Usually, 100 decision trees are used as the classifier's initial estimates (n_estimators=100), however this may be adjusted as needed. The classifier is trained using the training set of data using the fit() approach.The accuracy of predictions generated using the predict() approach based on test data is evaluated using the confusion matrix and the scikit-learn accuracy_score() function.Using a set of hyper parameters (n_estimators=500, max_depth=20, and max_leaf_nodes=10000), a specific random forest classifier is generated, trained on the training data, and then used to make predictions on the test data. The accuracy_score() function and a confusion matrix are also used to evaluate these predictions. Figure 1 shows the feature plot from using random forest classifier.

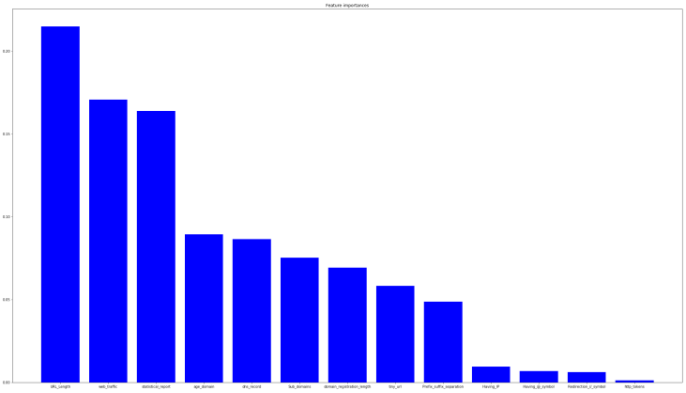

Fig. 1 Feature plot from using random forest classifier

IV. EXPECTED OUTCOMES

The present outcome is in line with expectations, as shown by a literature study and the features. The most suitable algorithms we could discover for our project were decision trees and random forests, as we noticed.

V. CONCLUSION

Phishing is a method for getting access to a user's private information via email or a website. Almost everything is now available online because so many people use it, whether it be for purchasing clothing, electronics, kitchenware, or for paying a phone, TV, or electricity bill. People are aware that using online techniques is preferable to standing in long queues. As a result, phishers have several chances to employ phishing scams. Despite extensive research in this area, there isn't a single method that can reliably detect every type of phishing attack. Attackers who use phishing always develop new methods as technology develops. As a result, we can find effective classifiers for phishing detection. The most important elements of a phishing attack defense are education and awareness. Internet consumers need to be aware of any professional security recommendations. Users should also be instructed not to carelessly click on links that take them to websites where they must enter sensitive information. Prior to viewing the website, it is essential to check the URL. The technology can be improved in the future to automatically determine whether a web page and an application are compatible with a certain browser. By adding a few extra elements, further work can be done to help identify between fraudulent and real websites. The security and dependability of the internet are gravely threatened by phishing, and phishing detection is a serious cause for concern. We looked at some of the shortcomings of the blacklist and heuristic evaluation approaches, two conventional methods for detecting phishing.

VI. REFERENCES

[1] Sopnil Nepal, Hemant Gurung, Roshan Nepal, “Phishing URL Detection Using CNN-LSTM and Random Forest Classifier”, a National College of Engineering, in 2022, doi: https://doi.org/10.21203/rs.3.rs-2043842/v2 .

[2] Challa Sai Bhanu Teja, Tanikella Sai Siva Sasank, Yakasiri Jeevan Sreeram Reddy, “Phishing Website Detection using DIFFERENT Machine Learning Techniques”, International Research Journal of Engineering and Technology, Vol.07(10), in 2020, e-ISSN: 2395-0056, p- ISSN: 2395-0072.

[3] Ammar Odeh, Ibrahim Abualhaol, smail Keshta, Ahmad Abushakra, “Phishing Website Detection Using Multilayer Perceptron”, International Journal of Information and Decision Sciences, Vol.24(6), in 2021.

[4] Abdulhamit Subasi, Emir Kremic, “Comparison of Adaboost with MultiBoosting for Phishing Website Detection”, Procedia Computer Science 168 (2020) 272–278.

[5] Said Salloum, Tarek Gaber, Sunil Vadera, Khaled Shaalan, “Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey”, Procedia Computer Science, Vol.189, 2021, pp.19-28.

[6] Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, “Phishing Detection Using Machine Learning Techniques”, 20 Sep 2020, arXiv:2009.11116v1 [cs.CR].

[7] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri, “Machine learning based phishing detection from URLs”, Expert Systems with Applications, Vol.117, 2019, pp.345-357.

[8] Ping Yi. Yuxiang Guan, Futai Zou, Yao, Wei Wang, Ting Zhu, “Web Phishing Detection Using a Deep Learning Framework”, Hindawi Wireless Communications and MobileComputing,vol.2018,2018,doi: https://doi.org/10.1155/2018/4678746, .

[9] Adwan Yasin, Abdelmunem Abuhasan, “An Intelligent Classification Model For Phishing Email Detection”, International Journal of Network Security & Its Applications,Vol.08, No.4, 2016.

[10] Adam Kavon Ghazi-Tehrani , Henry N. Pontell, “Phishing Evolves: Analyzing the Enduring Cybercrime”, Victims & Offenders An International Journal of Evidence- based Research, Policy and Practice, Vol.16(3), 2021.

[11] A. J. Burns, M. Eric Johnson, Deanna D. Caputo, “Spear phishing in a barrel: Insights from a targeted phishing campaign”, Journal of Organizational Computing and Electronic Commerce, Vol.29(1), 2019, ISSN: 1091-9392, pp.24-39.

[12] Gururaj Harinahalli Lokesh, Goutham BoreGowda, “Phishing website detection based on effective machine learning approach”, Journal of Cyber Security Technology, 2020,doi: https://doi.org/10.1080/23742917.2020.1813396.

[13] ShymalaGowri Selvaganapathy, Mathappan Nivaashini & HemaPriya Natarajan, “Deep belief network based detection and categorization of malicious URLs”, Information Security Journal: A Global Perspective, 2018, ISSN:1939-3555,doi: https://doi.org/10.1080/19393555.2018.1456577.

[14] Sonali Sharma, Prof. Somesh Kumar Dewangan, “Data Encryption Technique Using Random Number and Salective Encryption Algorithem”, International Journal of Engineering Research & Technology, Vol. 3(4), 2014, ISSN: 2278-0181.

[15] Yong Fang , Cheng Zhang, Cheng Huang , Liang Liu, Yue Yang, “Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism”,IEEEAccess,2019,doi:10.1109/ACCESS.2019 .2913705.

[16] Rabab Alayham Abbas Helmi, Chua Shang Ren, Arshad Jamal, Muhammad Irsyad Abdullah,” Email Anti-Phishing Detection Application”, 2019 IEEE 9th International Conference on System Engineering and Technology, 2019.

[17] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, Cleotilde Gonzalez,"Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol.63(1), 2019, pp.453-457,doi: https://doi.org/10.1177/1071181319631355.

[18] Andronicus A. Akinyelu, Aderemi O. Adewumi,"Classification of Phishing Email Using Random Forest Machine Learning Technique",School of Mathematics,Statistics and Computer Science,University ofKwaZulu-Natal,Vol.2014,doi: https://doi.org/10.1155/2014/425731.

[19] Hikmat Ullah Khan,Nazish Yousaf,Farah Aslam,Almas Anjum,Maryam Hamdani,"Phishing web site detection using diverse machine learning algorithms",The Electronic Library,Vol.38(1),2020.

[20] Ashina Sadiq,Muhammad Anwar,Rizwan A. Butt,Farhan Masud,MuhammadK.ShahzadShahid Naseem,Muhammad Younas, "A review of phishing attacks and countermeasures for internet of things-based smart business applications in industry 4.0", Human Behavior and Emerging Technologies, 3(2), 2021.

[21] Rana Alabdan,"Phishing Attacks Survey: Types, Vectors, and Technical Approaches",Security and Privacy in Social Networks and Solutions,Vol.12(10), No.168, 2020, doi: https://doi.org/10.3390/fi12100168.