INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, June 2023, pp : 140-143

e-ISSN : 2583-1062

Impact Factor : 5.725

DETECTING PHISHING WEBSITE USING MACHINE LEARNING

Shantanu Agnihotri[1], Dr. Santosh Kr. Dwivedi[2], Mr. Raghvendra Singh[3]

[1]UG Student of Department of Bachelor of Computer Application, Shri Ramswaroop Memorial College of Management, Lucknow, Uttar Pradesh, India

[2]Professor, Head of Department of Bachelor of Computer Application, Shri Ramswaroop Memorial College of Management, Lucknow, Uttar Pradesh, India

[3]Professor, Department of Bachelor of Computer Application, Shri Ramswaroop Memorial College of Management, Lucknow, Uttar Pradesh, India

ABSTRACT

Phishing attacks pose a significant and growing threat in the cyber world, resulting in substantial financial losses for internet users. These criminal activities employ various social engineering tactics to deceive users and obtain sensitive information. Phishing attempts can occur through different communication channels, including emails, instant chats, pop-up messages, and web pages. This research focuses on the development of a model capable of predicting the legitimacy of URL links, specifically distinguishing between legitimate and phishing URLs. To train the classification model, a diverse dataset was compiled from two sources: "Phish Tank," an open-source service that provides phishing URLs in multiple formats like CSV and JSON, and the University of New Brunswick dataset bank, which includes benign, spam, phishing, malware, and defacement URLs. The dataset consists of over 5,000 URLs, randomly selected and divided into 80,000 training samples and 20,000 testing samples. These samples are equally distributed between phishing and legitimate URLs. The classification process involves utilizing six different machine learning models and deep neural network algorithms to detect phishing URLs. By leveraging the features extracted from the URLs, such as address bar-based features, domain-based features, and HTML & JavaScript-based features, the model can accurately identify and differentiate between legitimate and phishing URLs. The primary objective of this study is to develop a web application software capable of effectively detecting phishing URLs. By integrating the developed model into the application, both individuals and companies can benefit from enhanced security measures. The application will enable users to authenticate any supplied link, ensuring its validity and assisting in identifying potential phishing attacks. In conclusion, this study contributes to the field of cybersecurity by providing a model for classifying URLs as either phishing or legitimate. The model's accuracy and effectiveness in identifying phishing attempts will prove invaluable in helping individuals and organizations combat these malicious activities. By verifying the authenticity of links, users can significantly enhance their ability to protect themselves against phishing attacks and safeguard their sensitive information.

1. INTRODUCTION

The Internet has become an important part of our lives for gathering and disseminating information, particularly through social media. According to Pamela (2021), the Internet is a network of computers containing valuable data, so there are many security mechanisms in place to protect that data, but there is a weak link: the human. When a user freely gives away their data or access to their computer, security mechanisms have a much more difficult time protecting their data and devices. Therefore, Imperva (2021) defines social engineering (a type of attack used to steal user data, including login credentials and credit card numbers) as a type of attack that is one of the most common social engineering attacks. The attack happens when an attacker fools a victim into opening an email, instant message, or text message as if it were from a trusted source. Upon clicking the link, the recipient is fooled into believing that they've received a gift and unsuspectingly clicks a malicious link, resulting in the installation of malware, the freezing of the system as part of a ransomware attack, or the disclosure of sensitive information. Phishing is a type of cyber attack that involves the use of fake websites or emails to trick people into providing sensitive information such as usernames, passwords, and credit card details. Phishing attacks are becoming increasingly sophisticated and difficult to detect, making them a significant threat to individuals, businesses, and organizations. Machine learning (ML) is a powerful tool for detecting and preventing phishing attacks. ML algorithms can analyze large amounts of data to identify patterns and anomalies that may indicate the presence of a phishing website. By training a machine learning model on a dataset of known phishing websites, the model can learn to recognize the characteristics of phishing websites and predict whether a new website is likely to be a phishing site. The goal of the project "Detecting phishing websites using ML" is to develop an ML model that can accurately detect phishing websites. This involves collecting a large dataset of known phishing websites, extracting features that are indicative of phishing, and training an ML algorithm to classify websites as either legitimate or phishing. The model can then be tested on a separate dataset of websites to evaluate its performance and refine the model as necessary. The project has significant implications for improving the security of online users and businesses. By detecting and preventing phishing attacks, the ML model can help to protect sensitive information and prevent financial losses due to cybercrime.

2. WORKFLOW

An extensive review was done on related topics and existing documented materials such as journals, e-books, and websites containing related information gathered which was examined and reviewed to retrieve essential data to better understand and know how to help improve the system.The methodology used to achieve the earlier stated objectives is explained below. The dataset collection consists of phishing and legitimate URLs which were obtained from open-source platforms. The dataset was then pre-processed that is cleaned up from any abnormality such as missing data to avoid data imbalance. Afterward, expository data analysis was done on the dataset to explore and summarize the dataset. Once the dataset was free from all anomalies, website content-based features were extracted from the dataset to get accurate features to train and test the model. An extensive review was done on existing works of literature and machine learning models on detecting phishing websites to best decide the classification models to solve the problem of detecting phishing websites. Hence, Series of these machine learning classification models such as Decision Tree, Support Vector Machine, XGBooster, Multilayer perceptions, Auto encoder Neural Network and Random Forest was deployed on the dataset to distinguish between phishing and legitimate URLs. The best model with high training accuracy out of all the deployed models was selected then integrated into a developed web application. Thus, a user can enter a URL link on the web application to predict if it is phishing or legitimate.

3. PROPOSED SYSTEM

The proposed phishing detection system utilizes machine learning models and deep neural networks. The system comprises two major parts, which are the machine learning models and a web application. These models consist of Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest. These models are selected after different comparison-based performances of multiple machine learning algorithms. Each of these models is trained and tested on a website content-based feature, extracted from both phishing and legitimate dataset. Hence, the model with the highest accuracy is selected and integrated into a web application that will enable a user to predict if a URL link is phishing or legitimate.

Benefits of the new system:

i. Will be able to differentiate between phishing(0) and legitimate(1) URLs

ii. It Will help reduce phishing data breaches for an organization

iii. It Will be helpful to individuals and organizations

iv. It is easy to use.

4. MODEL DEVELOPMENT

IThe model development method takes several models, tests them, and adds them to an iterative process until a model that meets the required requirements is developed. Figure 3.1 shows the steps used in the development of machine learning models using both supervised and unsupervised learning.The following are the stages to machine learning model development for phishing detection systems:

i. Data Collection- The data used to generate the datasets on which the models are trained are gotten from different open-source platforms. The dataset collection consists of phishing and legitimate URL dataset.The set of phishing URLs are collected from an open-source service called Phish Tank. This service provides a set of phishing URLs in multiple formats like CSV, JSON and so on that gets updated hourly. This dataset isaccessible from the phishtank.com website. From this dataset, over 5000 random phishing URLs are collected to train the ML models.The set of legitimate URLs are obtained from the open datasets of the University of New Brunswick, This dataset is accessible on the university website. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign URL dataset is considered for this project. From this dataset, Over 5000 random legitimate URLs are collected to train the ML models.

ii. Preprocessing- Data preprocessing is the first and crucial step after data collection. The raw dataset obtained for phishing detection was prepared by removing redundant and irregular data and also encoded using the One-Hot Encoding technique into a useful and efficient format suitable for the machine learning model.

iii. Exploratory data analysis- Exploratory data analysis (EDA) technique was used on the dataset after series of data cleaning. The data visualization method was employed to analyze, explore and summarize the dataset. These visualization consist of heat-map, histograms, box plots, scatter plots, and pair plots to uncover patterns and insights within data.

iv. Feature Extraction-Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones. Thus, Website content-based features were extracted from phishing and legitimate datasets such as the Address bar-based feature which consists of 9 features, Domain-based feature which consists of 4 features, and Html & JavaScript-based Feature which consists of 4 features. So, altogether 17 features were extracted for phishing detection.

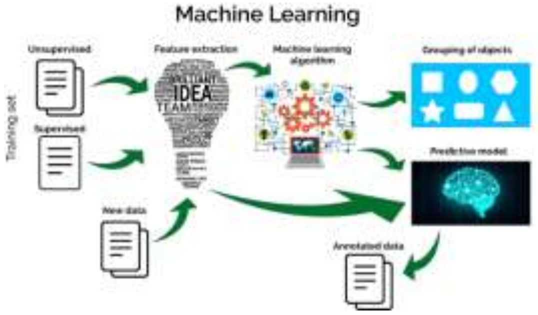

Figure 1: Machine Learning development process

5. ANALYSIS

The image as shown in figure 2 shows the distribution plot of how legitimate and phishing datasets are distributed base on the features selected and how they are related to each other. In figure 3 shows the plot of a correlation heat-map of the dataset. The plot shows correlation between different variables in the dataset.

In figure 4 , it shows the feature importance in the model for Decision tree classifier.

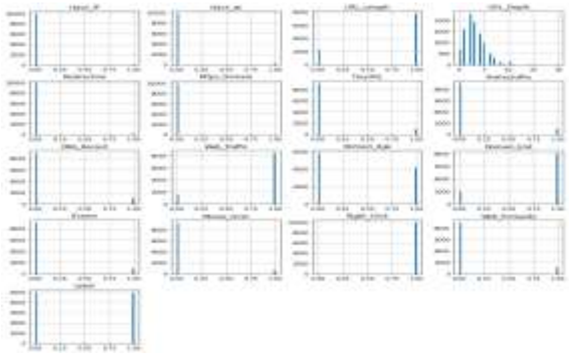

Figure 2 : Distribution plot of dataset base on the features selected

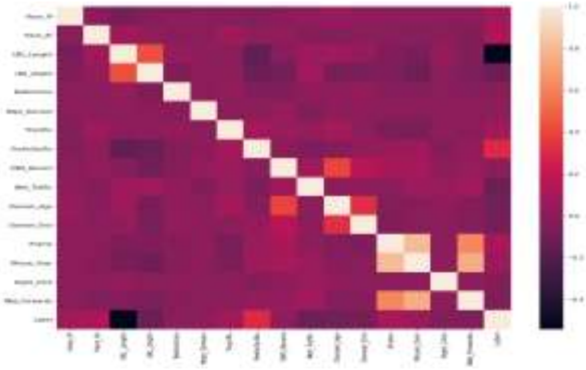

Figure 3: Correlation heat map of the dataset

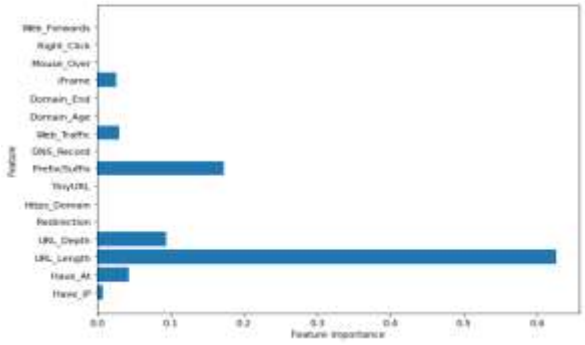

Figure 4: Feature importance for Decision Tree classifier

6. SCOPE OF WORK

This study explores data science and machine learning models that use datasets gotten from open-source platforms to analyze website links and distinguish between phishing and legitimate URL links. The model will be integrated into a web application, allowing a user to predict if a URL link is legitimate or phishing. This online application is compatible with a variety of browsers.

7. CONCLUSION

The system developed detects if a URL link is phishing or legitimate by using machine learning models and deep neural network algorithms. The feature extraction and the models used on the dataset helped to uniquely identify phishing URLs and also the performance accuracy of the models used. It is also surprisingly accurate at detecting the genuineness of a URL link.Through this project, one can know a lot about phishing attacks and how to prevent them. This project can be taken further by creating a browser extension that can be installed on any web browser to detect phishing URL Links.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to all those who have supported and guided me throughout the development of the "Detecting Phishing Website using ML" web application. First and foremost, I would like to extend my heartfelt thanks to Mr. Raghvendra Singh, my project guide, for his invaluable guidance, expertise, and constant encouragement. His profound knowledge and insights have been instrumental in shaping this project and ensuring its successful completion. I am sincerely grateful for his patience, mentorship, and unwavering support throughout this journey. I would also like to extend my thanks to the faculty and staff of SRMCM, Lucknow, for providing me with the necessary resources and a conducive learning environment to pursue this project. Their continuous support, both academically and administratively, has been crucial in enabling me to undertake this endeavor. Furthermore, I am indebted to all the farmers and consumers who participated in the development and testing of the "Detecting Phishing Website using ML" web application. Their valuable feedback and inputs have been instrumental in refining the platform and ensuring its user- friendliness and effectiveness. Lastly, I would like to acknowledge the contributions of all the individuals and organizations who have directly or indirectly supported this project. Their involvement and collaboration have been instrumental in shaping the vision and objectives of the "Detecting Phishing Website using ML" web application.

8. REFERENCES

[1] Abdelhamid, N., Thabtah F., & Abdel-Jaber, H. Phishing detection: A recent intelligent machine learning comparison based on models’ content and features," 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2017, pp. 72-77, DOI: 10.1109/ISI.2017.8004877.

[2] Anjum N. S., Antesar M. S., & Hossain M.A. (2016). A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps. Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Chengdu, China, 2016, pp. 9-15, DOI: 10.1109/SKIMA.2016.7916190.

[3] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques, Proceedings of IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 2070–2090.

[4] Ashritha, J. R., Chaithra, K., Mangala, K., & Deekshitha, S. (2019). A Review Paper on Detection of Phishing Websites using Machine Learning.Proceedings of International Journal of Engineering Research & Technology (IJERT), 7, 2. Retrieved from www.ijert.org.

[5] Anti-Phishing Working Group (APWG) Phishing activity trends report the first quarter. (2014) Retrieved from http://docs.apwg.org/reports/apwg trends report q1 2014.pdfAPWG report. (2014). Retrieved from http://apwg.org/download/document/245/APWG Global Phishing Report 2H 2014.pdf 110

[6] Ayush, P. (2019). Workflow of a Machine Learning project. Retrieved from https://towardsdatascience.com/workflow-of-a-machine-learning-projectec1dba419b94