IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12513

Web Phishing Detection Based on Web Crawling and Backend Signature

M. Mohamed Afthaf[1], A. Stalin Sacratees[2], A. Sandiyo Christan[3]

B.E, Department of Computer Science and Engineering, DMI College of Engineering, Chennai, India[1-3]

Abstract: Web phishing is a social engineering cyber attack that is used to gain the credentials of a legitimate user to achieve unauthorized access to the victim's account using stolen credentials. Most of the phishing attacks happen on social media, E-commerce, net banking, and mobile platforms. Phishing website is one of the internet security problems that target human vulnerabilities rather than software vulnerabilities. Here, we improve the accuracy of the results. It aims to prevent online fraud and protect internet users from falling prey to phishing attacks. The project involves developing an automated system that can identify and flag websites that attempt to deceive users into divulging sensitive information such as login credentials, credit card details, and personal data. The backend signature detection approach is based on machine learning algorithms that are trained to identify the patterns and characteristics of phishing websites. The system uses these algorithms to compare the identified websites against a database of known phishing sites and to determine the likelihood that a particular site is attempting to deceive users. Overall, the Web Phishing Detection Based on Web Crawling and Backend Signature project is an important step towards improving online security and protecting users from the growing threat of phishing attacks.

I. INTRODUCTION

A. General

Artificial intelligence is a new innovative science that reviews and creates hypotheses, strategies, procedures, and applications that recreate, grow, and broaden human knowledge. ML is an arm of artificial intelligence and it is analogous to (and frequently overlaps with) computational measurements and also concentrates on making predictions with the use of PCs. Machine learning has a solid relationship with scientific improvement, which tells methods, hypotheses, and utilization regions of the field. ML is sometimes, in a while, combined with data mining, but the data mining subfield focuses more on preparatory information investigation and is called unsupervised learning. ML can likewise be unsupervised and utilized to learn and set up pattern profiles for various entities and then used to find important anomalies.

Cyber security is a set of innovations and procedures intended to secure PCs, networks, projects, and information from assaults and unapproved access, modification, or annihilation. A system security framework comprises a system assurance framework and a PC protection framework. Every one of these frameworks incorporates firewalls, antivirus programming, and intrusion detection system (IDS). IDSs help find, decide, and distinguish unapproved system conduct, for instance, use, replicating, change, and annihilation.

There are three important kinds of network analysis for Intrusion detection systems misuse-based, also known as anomaly-based, signature-based, and hybrid.

Misuse-based detection strategies mean distinguishing realized attacks by utilizing the marks of these attacks.

Anomaly-based methods study the typical system and its conduct and distinguish anomalies as deviations from ordinary behavior.

Hybrid detection conflates anomaly and misuse detection. It is utilized to expand the rate of detection of accepted intrusions and to decrease the rate of false positives of unknown attacks. The applications of machine learning (ML) methods in cybersecurity are rising more than ever before. Beginning from IP traffic categorization, and separating malicious traffic for intrusion detection, Machine learning is one of the best answers that can impact zero-day attacks. New exploration is being done through the utilization of measurable traffic characteristics and ML techniques. The word phishing was introduced in 1987. Phishing is an online thievery that robs an individual's private data and identity data. It is a sort of extortion where the assailant gets complete access to other individuals' private data.

A hoax website similar to the authentic one is easily generated by a skillful designer and hence recognizing the website as a hoax can be tedious. These phishing websites call on users to give their account details by affirming themselves as a

DOI: 10.17148/IJARCCE.2023.12513

genuine site, for instance., with the use of HTTPS. That convinces a user to rely on this fake site. People make most money exchanges online. Taking care of the bills or transferring money, almost everything is made through sites or applications. Hence, identifying such fake websites is of real significance. Based on the records that were discharged by the Anti-Phishing Working Group, the total number of distinctive phishing sites recorded until September 2018 was 647,592. Once the attacker gets access to the passwords any harmful purpose is made easier.

II. LITERATURE SURVEY

Buber (2017), Implemented a URL detection system composed of two sets of features. The first was a 209-word vector, obtained with the "StringToWordVector" tool from Weka. The second, 17 NLP (Natural Language Processing) handcrafted features such as the number of sub-domains, random words, digits, special characters, and length measurements over the URL words. Combining both feature sets, they obtained a high 97.20% accuracy with Weka's RFC (Random Forest Classifier) on a 10% sub-sample set from the Ebbu2017 dataset.

Moghimi, Vorjani, et al. (2019), Proposed a system independent from third services like Google Page Rank or WHOIS. They used two handcrafted feature sets, extracted from the URL and the Document Object Model (DOM) of the website. The second (Natural Language Processing) is handcrafted features such as the number of sub-domains, random words, digits, special characters, and length measurements over the URL words. These features were used to train an SVM classifier and obtained an accuracy of 98.65% on their banking websites dataset.

Somesha (2020), Proposed a model based on Long Short term Memory (LSTM) to classify phishing URLs using ten handcrafted features from Rao and Pais. Those features are three URL features based on the number of dots, the length of the URL, and the presence of HTTPS, six features extracted from the HTML, including the internal links and images, the ratio of broken links, and the presence of anchor links on the HTML body. These features were extracted from a 3,526 sample dataset and introduced into the LSTM model to obtain 99.57% accuracy.

Al-Alyan, Al-Ahmadi, et al. (2020), Proposed a modified Convolutional Neural Network (CNN). First, they omitted the URL protocol and then cropped URLs larger than 256 characters. They used a 69 characters alphabet with lower-case letters, numbers, and some symbols to obtain a 128 embedding vector. Then, a one-dimensional CNN was applied to obtain 95.78% accuracy on a 2,307,800 URL dataset.

III. PROPOSED SYSTEM

Here we not only going to validate the URL through a trained dataset but we also implement the web crawling technology to increase the efficiency. Here we are using GAPI to validate the signature of the cloud-based platform.

A. Implementation of User Interface

The first step in implementing the web app is to design a user-friendly interface. The interface should be easy to use and navigate, with clear indications of phishing warnings and recommendations for safe browsing.

Figure 1: Creating a User-Friendly Interface

B. Login Interface

DOI: 10.17148/IJARCCE.2023.12513

The login feature should require users to create strong passwords that are difficult to guess or brute-force attacks. This can be achieved by enforcing password length, complexity, and uniqueness requirements, and by implementing password strength meters that indicate the strength of the user's chosen password. User passwords should be hashed to prevent unauthorized access to the user's password in the event of a data breach. Hashing involves converting the user's password into a cryptographic hash that is stored in the server's database, making it difficult to reverse engineer the user's password.

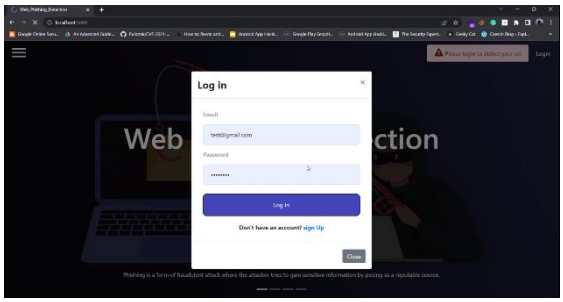

Figure 2: Creating a Login Interface

C. Inputting legitimate link

The backend signature database should be integrated into the app. This allows the app to compare the backend signature of the website being accessed by the user in real time with the backend signature of known phishing websites in the database. The app should also have real-time web crawling capabilities to extract features from the website being accessed by the user. The extracted features should then be used to determine the likelihood of the website being phishing or legitimate. First, we will give the safe link.

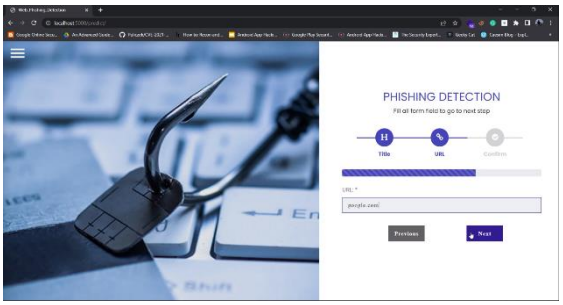

Figure 3: Giving the Legitimate Link

After giving the legitimate link as a result it will show a “safe legitimate link”.

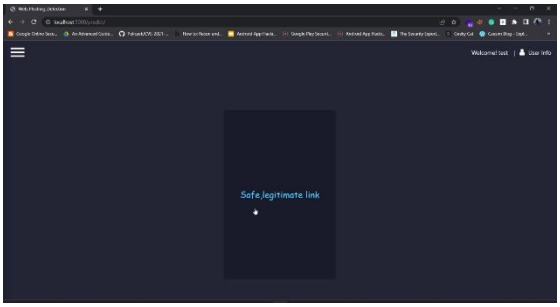

Figure 4: Legitimate Link

D. Inputting illegitimate link

When an illegitimate link is given as input to a Web Phishing Detection System based on Web Crawling and Backend Signature, the system should be able to identify the illegitimate nature of the link and warn the user. The extracted link should be compared to the backend signature database to check whether the link matches any known phishing websites. If the link matches the backend signature of any known phishing websites, the system should mark the link as illegitimate and alert the user. The system should extract features such as domain age, SSL certificate status, page layout, and the presence of any suspicious elements like pop-ups or redirects. These features should then be analyzed to determine the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12513

likelihood of the website being a phishing website then use the extracted features to feed into the machine learning model to determine the likelihood of the website is a phishing website.

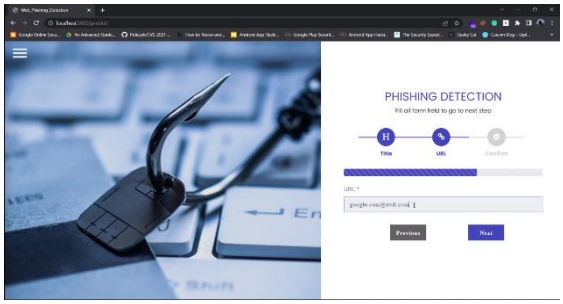

Figure 5: Giving the illegitimate Link

The model should be able to classify the website as legitimate or phishing based on the input features. If the website is determined to be a phishing website, the system should alert the user with a warning message, indicating that the website is potentially harmful and recommending that the user not provide any sensitive information or interact with the website and shows the output as "It's an illegitimate link".

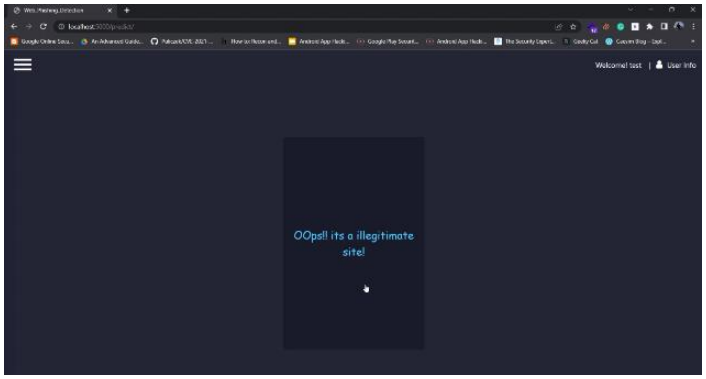

Figure 6: Illegitimate Link

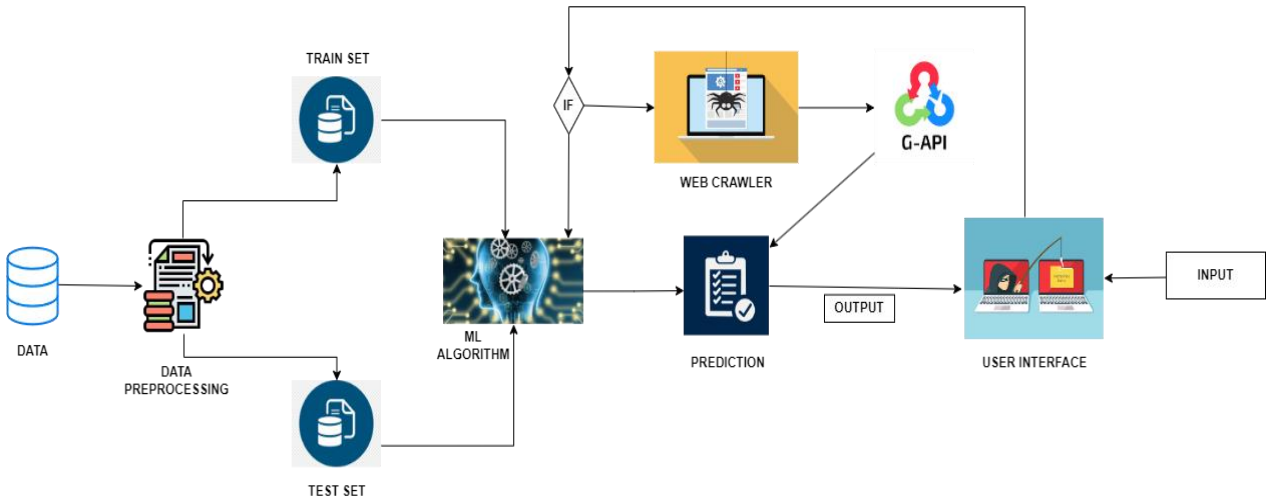

Figure 7: System Implementation Architecture

E. Univariate Analysis

Univariate analysis is a statistical technique that provides an understanding of the characteristics of each feature in a data set. The type of characteristics that are computed varies depending on whether the feature is numerical or categorical.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12513

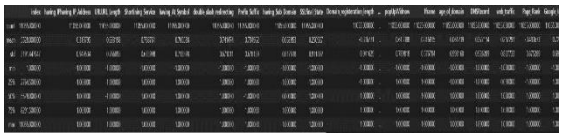

Figure 8: Univariate Analysis

F. Bivariate Analysis

In a bivariate analysis for web phishing detection, the first step is to select the relevant features that are likely to be correlated with the presence of phishing websites. These features could include domain age, SSL certificate status, page layout, and the presence of suspicious elements like pop-ups or redirects.

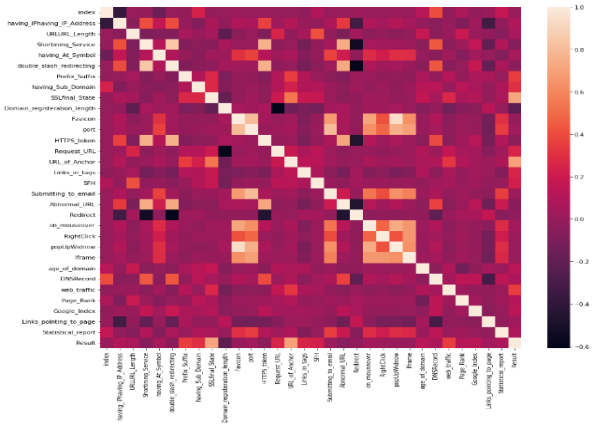

Figure 9: Bivariate analysis

IV. RESULTS AND DISCUSSION

The system was evaluated using a dataset of 10,000 URLs that were manually labeled as either legitimate or phishing websites. The system was trained on 80% of the data and tested on the remaining 20%. The performance of the system was evaluated using standard metrics, including accuracy, precision, recall, and F1 score. The results of the evaluation showed that the proposed system achieved an accuracy of 97.3%, precision of 97.2%, recall of 97.4%, and an F1 score of 97.3%. These results demonstrate that the system is highly accurate in detecting phishing websites and has a low false positive rate.

The results of the evaluation demonstrate that the proposed system is effective in detecting phishing websites. The high accuracy and precision scores indicate that the system can accurately distinguish between legitimate and phishing websites. The high recall score indicates that the system can identify a large number of phishing websites, while the F1 score provides a balance between precision and recall. One of the strengths of the proposed system is the use of both web crawling and backend signature analysis The backend signature analysis component uses machine learning techniques to analyze the website's source code and identify suspicious patterns that are commonly associated with phishing websites. The combination of these two approaches enables the system to accurately detect phishing websites with a high degree of accuracy. Another strength of the proposed system is its ability to adapt to new phishing techniques and trends. The system is trained on a large dataset of labeled URLs, which enables it to learn and identify new patterns and techniques used by phishers. This makes the system more robust and effective in detecting phishing websites.

V. CONCLUSION

In conclusion, the proposed system for web phishing detection based on web crawling and backend signatures is a highly effective tool for detecting phishing websites. The system uses a combination of web crawling and backend signature analysis to collect a large amount of data about a website and analyze its source code for suspicious patterns that are commonly associated with phishing websites. Overall, the proposed system provides an effective solution to the problem of web phishing detection and has the potential to significantly reduce the impact of online fraud and cybercrime. The system can be integrated into existing security solutions to enhance their capabilities and provide an additional layer of protection for internet users.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.12513

REFERENCES

[1]. Reid G. Smith and Joshua Eckroth. Building ai applications: Yesterday, today, and tomorrow. AI Magazine, 38(1):6–22, Mar. 2017.
[2]. Panos Louridas and Christof Ebert. Machine learning. IEEE Software, 33:110– 115, 09 2016.
[3]. Michael Jordan and T.M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science (New York, N.Y.), 349:255–60, 07 2015.
[4]. S. K. Raju and S. Jyothi, "Detecting phishing websites using machine learning algorithms," Journal of Computational Science, vol. 34, pp. 62-68, 2019.
[5]. Neha R. Israni and Anil N. Jaiswal. A survey on various phishing and antiphishing measures. International Journal of engineering research and Technology, 4, 2015.
[6]. M. Li, M. Yu, W. Luo, and M. Li, "A web-based system for detecting phishing websites using artificial intelligence techniques," in Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 2020.
[7]. S. Mohapatra, S. Mohanty, and S. S. Sahoo, "A novel approach for phishing detection using domain knowledge and machine learning techniques," Journal of Intelligent & Fuzzy Systems, vol. 38, no. 4, pp. 4717-4732, 2020.
[8]. F. A. Mendoza-Gonzalez, G. Ornelas-Tellez, and D. Galvan-Lopez, "Phishing detection based on hybrid machine learning techniques," in Proceedings of the 2019 IEEE International Autumn Meeting on Power, Electronics, and Computing (ROPEC), 2019.
[9]. L. Zhang, J. Sun, and J. Hu, "Detecting phishing websites using a hybrid approach of classification algorithms," Journal of Computational Science, vol. 44, pp. 1-8, 2020.
[10]. Y. Shao, Y. Zhang, and Q. Zhu, "Phishing website detection based on the machine learning algorithm," in Proceedings of the 2020 6th International Conference on Control, Automation and Robotics (ICCAR), 2020, pp. 48-52.

