IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Review of Prediction of Phishing Websites using Machine Learning Approaches

Miss. Saumya Jain[1], Prof. Chetan Gupta[2]
M.Tech. Research Scholar[1], Assistant Professor[2]
Department of Computer Science Engineering[1, 2]
SIRTS, Bhopal, India[1], SIRT, Bhopal, India[2]

Abstract: Phishing is known as one of the oldest forms of Cyber attacks. A phishing website is a domain similar in name and appearance to an official website. They're made in order to fool someone into believing it is legitimate. Today, phishing schemes have gotten more varied, and are potentially more dangerous than before. Artificial intelligence based machine and deep learning techniques is capable to predict the phishing websites. This paper presents the review of prediction of phishing websites using machine learning approaches.

***Index Terms* - Phishing Websites, Machine Learning, Accuracy, Error rate.**

I. INTRODUCTION

Phishing is the process of attempting to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a trustworthy entity using bulk email which tries to evade spam filters. Emails claiming to be from popular social web sites, banks, auction sites, or IT administrators are commonly used to lure the unsuspecting public. It’s a form of criminally fraudulent social engineering. Now day’s digital operations became more important, and people started to depend on new initiatives such as the cloud and mobile infrastructure. Consequently, the number of cyber attacks such as phishing has increased. Phishing websites can be detected using machine learning by classifying the websites into legitimate or illegitimate websites [1].

Now-a-days there are different types of cybercrime, Phishing is one of cyber-attacks where attackers impersonate as a member of legitimate institutions or organizations through an email, text message, advertisements or through any means to steal sensitive information which results to loss of personal and sensitive information such as account no, social security no, credit card no etc. Phishing attack has been increasing exponentially. In this attack mostly innocent users are comes to losses their sensitive, unique, personal, valuable and secure data and information’s. Many hackers are accomplished through phishing attacks where client are trapped into interacting with web-pages which looks like to be legitimate websites [2]. Enormous network connectivity, big data, the Internet of Things (IoT), digitalization of the world, and the use of social websites and apps has brought enormous institutional and individual security challenges. The conventional security system often fails to provide cyber security to institutions and individuals. Artificial Intelligence (AI) is highly adaptive and smart to handle the volatile cyber security environment. AI plays a prudent role in access control, user authentication and behavior analysis, spam, malware, and bonnet detection [3].

Scraping bots or scrapers are computer programs that automatically fetch data from the Internet. The price of products is illegally copied from different e-commerce stores and pasted into their sites for their benefit. Various web traffic surveys show that automated programs account for approximately half of all website traffic. However, an E-commerce website concerns many security issues that remain unresolved despite a significant increase in E-commerce development. One of the most damaging attacks on E-commerce sites is the price scraping when it comes to competitors [4].

Third-party tracking on the Web has been used for collecting and correlating user's browsing behavior. Due to the increasing use of ad-blocking and third-party tracking protections, tracking providers introduced a new technique called CNAME cloaking. It misleads Web browsers into believing that a request for a sub domain of the visited website originates from this particular website, while this sub domain uses a CNAME to resolve to a tracking-related third-party domain. This technique thus circumvents the third-party targeting privacy protections. The goals of this paper are to characterize, detect, and protect the end-user against CNAME cloaking based tracking. Firstly, we characterize CNAME cloaking-based tracking by crawling top pages of the Alexa Top 300,000 sites and analyzing the usage of CNAME cloaking with CNAME blacklist, including websites and tracking providers using this technique to track users' activities [6].

Artificial intelligence (AI) in web development is a new sector that a lot of people are into recently. AI continues to evolve and grow, and plays an increasingly important role in the web app development space [1]. When it comes to developing innovative and more sophisticated web applications, the involved technologies continue to play a bigger role. With the involvement of the internet into our daily lives, particularly businesses are enjoying the aspects of AI. Precisely, companies use AI in proper marketing of their products and enhancing their brand visibility by building their websites and web applications [2]. AI or Machine Learning (ML) models are able to help web app developers to solve problems related to security, user experience, content analysis, quality assurance and much more. This presents the need for a framework or tool that can allow third party developers to seamlessly build an AI based app [7].

User demand for blocking advertising and tracking online is large and growing. Existing tools, both deployed and described in research, have proven useful, but lack either the completeness or robustness needed for a general solution. Existing detection approaches generally focus on only one aspect of advertising or tracking (e.g. URL patterns, code structure), making existing approaches susceptible to evasion. In this work we present AdGraph, a novel graph-based machine learning approach for detecting advertising and tracking resources on the web. AdGraph differs from existing approaches by building a graph representation of the HTML structure, network requests, and JavaScript behavior of a webpage, and using this unique representation to train a classifier for identifying advertising and tracking resources [9].

Cyber security is a most important trepidation in the widespread adoption of internet technologies in the everyday activities of human being. Even though more sophisticated technologies emerged on the Internet, but different kinds of attacks and threats are also increasing day by day. Cyber attacks causes loss of customer confidence in adopting internet based applications. Phishing attack is one of the common vulnerabilities in the cyber space. Most of the anti-phishing solutions proposed so are focused only on a single issue and needs improvement. For malicious web page detection and prevention, an intelligent multi agent solution is proposed with the help of machine learning methods [10].

II. REVIEW OF LITERATURE

F. Yahya et al., [1] the purpose of the study is to conduct a mini-review of the existing techniques and implement experiments to detect whether a website is malicious or not. The dataset consists of 11,055 observations and 32 variables. Three supervised learning models are implemented in this study: Decision Tree, K-Nearest Neighbor (KNN), and Random Forest. The three algorithms are chosen because it provides a better understanding and more suitable for the dataset. Based on the experiments undertaken, the result shows Decision Tree has an accuracy of 91.16% which is the lowest compared to the other models, 97.6 % for the KNN model which is the highest among all the models and 94.44% accuracy for the Random Forest model. Through comparisons between the three models, KNN was the prime candidate for the best model considering that it has the highest accuracy. However, Random Forest is deemed more suitable for the dataset even though the accuracy is lesser because of the lowest false-negative value than the other models. The experiments can be further investigated with different datasets and models for comparative analysis.

K. S. Swarnalatha et al.,[2] Phishers are appears in many platform of communication such as in the form of VOIP, message and e-mails which is not real. Commonly users have many accounts on various websites including social media, email, and also in bank. So that innocent users are the most vulnerable targets for these types of attack. This happened because most of the peoples are unknown of the sensitive data, which helps them to get their information successfully. As of 2020, phishing is the most common attack performed by cyber criminals according to FBI's Internet Crime Complaint Centre. First phishing Datasets are collected from phish tank and then legitimate websites are collected from University of New Brunswick and then dataset is preprocessed using wrapper and filter method so that it covers the dataset which gets missed, tampered and unstructured.

S. M. Istiaque et al.,[3] Machine learning (ML) models are the building blocks of AI. In this research, a novel practical approach is followed to prove the effectiveness of AI in the field of cyber-security. Multiple machine learning algorithms are applied to prove that. A two-step suitability test is conducted in this study. In the 1st step, the KDD'99 data set is used to train and test the AI models. In the second step, train models are repeatedly tested on a fresh data set, NSL-KDD. Finally, the testing results are compared to prove the authenticity of AI in cyber-security. An absolute practical approach and reliable outputs of various AI models prove AI's suitability in cyber-security.

R. Yaqoob et al.,[4] aims to uncover the fact that both humans and bots are involved in stealing the prices of products from different e-commerce stores without the permission of original sites. This paper discusses the well-known price scraping tools and techniques. First, we have performed real-time price scraping attacks through custom XPATH, and we have obtained results which show that some platforms are still vulnerable for bots e.g., Alibaba and eBay. These sites do not restrict the copy of XPATH. Whereas, Amazon and daraz.pk blocks the price code from HTML document for the scraper to copy the custom XPATH. Moreover, we propose solutions to mitigate price scraping attacks on different e-commerce stores.

M. Min et al.,[5] The emergence of illegal online gambling (IOG) has led to an increase in gambling addiction and threats to cyber security. Since IOG is advertised through short message service (SMS), we propose a novel system to detect and extract uniform resource locator (URL) information from SMS spam. For majority of the cases, these URLs are not directly linked to a real website address, which causes difficulties for detection systems. In order to address this problem, we utilized a readable transformation technique (RTT). The resulting suggestions can be enhanced to extract URLs automatically from SMS spam, which will allow local authorities identify and block IOG operations.

H. Dao et al.,[6] privacy protection extensions are largely ineffective to deal with CNAME cloaking-based tracking except for Firefox with a developer's version of the uBlock Origin extension. Secondly, we propose a supervised machine learning-based approach to detect CNAME cloaking-based tracking without the on-demand DNS lookup. We show that the proposed approach outperforms well-known tracking filter lists. Finally, to circumvent the lack of DNS API in Chrome-based browsers, we design and implement a prototype of the supervised machine learning-based browser extension to detect and filter out CNAME cloaking tracking, called CNAMETracking Uncloaker. Our evaluation shows that CNAMETracking Uncloaker is able to filter out CNAME cloaking-based tracking requests without performance degradation when compared with the vanilla setting on the Chrome browser.

R. Nanjundappa et al.,[7] present an AI Enabled Web Contents Authoring Framework (AWAF), where AI models can be simply dragged into workspace, provide options to build, train Deep learning models using a simple web visual interface, and ultimately ship the AI features into the web application. Also, we provide an option to connect together smart blocks called AI Nodes, to create our custom deep learning models. These AI Nodes are designed flexible enough to reap the advantages of portability and reusability. And, laterally, we also focus on assigning or distributing the computational AI Nodes to capable IOT-edge devices like high-end TV etc. to leverage their hardware capabilities in order to increase the overall responsiveness of the AI application on low-end devices.

K. E. Aydın et al.,[8] Categorization of web sites is an important problem and has many practical applications. One such application is parental control for safe internet for children. Failure to classify websites by specific rules makes it difficult to access information, as well as leaving many users of different age groups with the harmful side of the Internet. Current secure internet solutions are not comprehensive or cannot be customized. Furthermore, the fact that the blocking orders issued by the courts do not cover all harmful sites and these websites change their domains so often. Thus, dynamic classification of websites using the text data is very important.

U. Iqbal et al.,[9] AdGraph considers many aspects of the context a network request takes place in, it is less susceptible to the single-factor evasion techniques that flummox existing approaches. We evaluate AdGraph on the Alexa top-10K websites, and find that it is highly accurate, able to replicate the labels of human-generated filter lists with 95.33% accuracy, and can even identify many mistakes in filter lists. We implement AdGraph as a modification to Chromium. AdGraph adds only minor overhead to page loading and execution, and is actually faster than stock Chromium on 42% of websites and AdBlock Plus on 78% of websites. Overall, we conclude that AdGraph is both accurate enough and perform enough for online use, breaking comparable or fewer websites than popular filter list based approaches.

N. Megha et al.,[10] The proposed approach detects both phishing sites and websites with malicious content. This multi-agent system contains four autonomous intelligent agents, which communicate with each other using the Extensible Messaging and Presence Protocol (XMPP) for decision-making. The first is a monitoring agent, second and third is for decision-making (using the machine-learning classifiers) and the fourth is for action-performing. The first agent is responsible for extracting URLs. It passes the extracted URLs to the second agent for feature extraction and classification. If any phishing is detected, the second agent communicates with the fourth agent and the site is blocked. Otherwise, the second agent communicates with the third agent for malicious script detection. If any malicious script is detected then the fourth agent blocks the entire web page. We have tested the performance and accuracy of the proposed method and obtained results ensures its efficiency.

S. S. Hashmi et al.,[11] Websites employ third-party ads and tracking services leveraging cookies and JavaScript code, to deliver ads and track users' behavior, causing privacy concerns. To limit online tracking and block advertisements, several ad-blocking (black) lists have been curated consisting of URLs and domains of well-known ads and tracking services. Using Internet Archive's Wayback Machine in this paper, we collect a retrospective view of the Web to analyze the evolution of ads and tracking services and evaluate the effectiveness of ad-blocking blacklists. We propose metrics to capture the efficacy of ad-blocking blacklists to investigate whether these blacklists have been reactive or proactive in tackling the online ad and tracking services. We introduce stability metric to measure the temporal changes in ads and tracking domains blocked by ad-blocking blacklists and diversity metric to measure the ratio of new ads and tracking domains detected.

T. Vo et al.,[12] There is an explosion in the advertisements over web nowadays. Most of the websites we visit contain ads, even Facebook, Google and Twitter. Sometimes, it could also appear that someone is spying on us because there are incidents like ads that show up with the content exactly what you have been searching not long ago. Such events happen as a result of Web Tracking. Initially, ads were meant to support businesses and companies to market their products and persuade the users to purchase them. Web Tracker were meant to track the user interaction with the website so it can improve the user experience. However, some of these have allowed ads as advertisements, which may take advantage of these functionalities to steal the user’s sensitive information.

After the literature survey following problem is identified in previous research work-

- Low accuracy rate of true data prediction from given Phishing Websites dataset.
- More classification error.
- Low precision and F measure value.

Objective-

- The objective of the proposed research work is to apply suitable machine learning technique to prediction of the phishing websites.
- Implementation will be performed using the python spyder 3.7 software.

III. PROPOSED STRATEGY

- Load the phishing websites dataset from the Kaggle [13].

In this step, the phishing websites dataset will be downloaded from kaggle source. It is a large dataset providing company. Then load this dataset into the python environment.

- Visualizing the Dataset

Now open the dataset files and view the various data in term of features like url, length_url, length_hostname, ip etc.

- Pre-process the Dataset

Now the data preprocess step applied, here data is finalize for processing. Missing data is either removal or replace form constant one or zero in this step.

- Splitting the Dataset into training and testing

In this step, the final preprocessed of dataset is divided into the training and the testing dataset. In the machine learning, firstly the machine is trained through given dataset then it comes in tested period for remaining dataset.

- Classification Using Machine Learning Algorithm

Now apply the machine learning technique to find the performance parameters. The existing work applied several techniques. In proposed method, we apply the SVM and KNN method and optimize the better results than other approach.

- Performance Metrics

(Accuracy, Precision, Recall, F1 - Score)

Now the performance parameters are calculated in terms of precision, recall, f-1 measure, accuracy etc by using the following formulas-

True Positive (TP): predicted true and event are positive.

True Negative (TN): Predicted true and event are negative.

False Positive (FP): predicted false and event are positive.

False Negative (FN): Predicted false and event are negative.

$$\begin{aligned}
Precision &= \frac{|TP|}{|TP| + |FP|} \\
Recall &= \frac{|TP|}{|TP| + |FN|} \\
F1 &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \\
Accuracy &= \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}
\end{aligned}$$

IV. CONCLUSION

The websites exactly seems to be semantically as well as visually to the original websites. The main idea of the phisher or hackers is to gain and purloin the critical information such as credential account, username, password and other private information related to any organization and company. According to phishing or web spoofing techniques is one examples of social engineering attack. This paper present the review of the previous work based on the phishing website detection. In the future take some suitable dataset from kaggel machine learning repository and apply the classification algorithm to achieve the better performance.

REFERENCES

1. F. Yahya et al., "Detection of Phishing Websites using Machine Learning Approaches," 2021 International Conference on Data Science and Its Applications (ICoDSA), 2021, pp. 40-47, doi: 10.1109/ICoDSA53588.2021.9617482.
2. K. S. Swarnalatha, K. C. Ramchandra, K. Ansari, L. Ojha and S. S. Sharma, "Real-Time Threat Intelligence-Block Phising Attacks," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1-6, doi: 10.1109/CSITSS54238.2021.9683237.
3. S. M. Istiaque, M. T. Tahmid, A. I. Khan, Z. A. Hassan and S. Waheed, "Artificial Intelligence Based Cybersecurity: Two-Step Suitability Test," 2021 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), 2021, pp. 1-6, doi: 10.1109/SOLI54607.2021.9672437.
4. R. Yaqoob, Sanaa, M. Haris, Samadyar and M. A. Shah, "The Price Scraping Bot Threat on E-commerce Store Using Custom XPATH Technique," 2021 26th International Conference on Automation and Computing (ICAC), 2021, pp. 1-6, doi: 10.23919/ICAC50006.2021.9594223.
5. M. Min, J. J. Lee, H. Park and K. Lee, "Honeypot System for Automatic Reporting of Illegal Online Gambling Sites Utilizing SMS Spam," 2021 World Automation Congress (WAC), 2021, pp. 180-185, doi: 10.23919/WAC50355.2021.9559478.
6. H. Dao, J. Mazel and K. Fukuda, "CNAME Cloaking-Based Tracking on the Web: Characterization, Detection, and Protection," in IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3873-3888, Sept. 2021, doi: 10.1109/TNSM.2021.3072874.
7. R. Nanjundappa et al., "AWAF: AI Enabled Web Contents Authoring Framework," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-5, doi: 10.1109/INDICON49873.2020.9342385.
8. K. E. Aydın and S. Baday, "Machine Learning for Web Content Classification," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, pp. 1-7, doi: 10.1109/ASYU50717.2020.9259833.
9. U. Iqbal, P. Snyder, S. Zhu, B. Livshits, Z. Qian and Z. Shafiq, "AdGraph: A Graph-Based Approach to Ad and Tracker Blocking," 2020 IEEE Symposium on Security and Privacy (SP), 2020, pp. 763-776, doi: 10.1109/SP40000.2020.00005.

10. N. Megha, K. R. Remesh Babu and E. Sherly, "An Intelligent System for Phishing Attack Detection and Prevention," 2019 International Conference on Communication and Electronics Systems (ICCES), 2019, pp. 1577-1582, doi: 10.1109/ICCES45898.2019.9002204.
11. S. S. Hashmi, M. Ikram and M. A. Kaafar, "A Longitudinal Analysis of Online Ad-Blocking Blacklists," 2019 IEEE 44th LCN Symposium on Emerging Topics in Networking (LCN Symposium), 2019, pp. 158-165, doi: 10.1109/LCNSymposium47956.2019.9000671.
12. T. Vo and C. Jaiswal, "ADREMOVER: THE IMPROVED MACHINE LEARNING APPROACH FOR BLOCKING ADS," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2019, pp. 1-4.
13. https://www.kaggle.com/datasets/isatish/phishing-dataset-uci-ml-csv?select=uci-ml-phishing-dataset.csv.