IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 15, May 2023

Employing Multi-Class Classification Techniques to Identify Harmful URLs

Mr. Vedant Rahane[1], Mr. Rushikesh Pokharkar[2], Mr. Aditya Wagh[3],Prof. K. U. Rahane[4]

Students, Department of Computer Engineering[1,2,3]
Professor, Department of Computer Engineering[4]
Amrutvahini College of Engineering, Sangamner, Maharashtra, India

Abstract*:* *The main method for hosting unsolicited content, such as spam, malicious ads, phishing, and drive-by exploits, to mention a few use of a malicious Uniform Resource Locator (URL), sometimes known as a malicious website. It is crucial to quickly identify the rogue URLs. The contemporary community of Internet users, which is becoming more integrated, is seriously threatened by malicious events. A popular method to find harmful events is detection based on network traffic features and machine learning techniques. Cybersecurity threats including ransomware, phishing, malware injection, etc. have significantly increased recently on many websites throughout the world. Numerous financial institutions, e-commerce businesses, and individuals suffered significant financial losses as a result. Since new attack types are being developed daily, controlling a cybersecurity attack in such a situation is a significant problem for cybersecurity specialists. Identifying dangerous URLs using lexical characteristics and a boosted tree-based machine learning strategy will be used in this project. Three well-known machine learning ensemble classifiers—Random Forest, Light GBM, and XGBoost—will be applied. We will be using a Malicious URLs dataset from Kaggle of 6,51,191 URLs, out of which 4,28,103 benign or safe URLs, 96,457 defacement URLs, 94,111 phishing URLs, and 32,520 malware URLs.*

Keywords: Malicious URL, Machine Learning, Cybersecurity, Detection of Fraudulent Traffic,Lexical Features, Multi-class Classification.

I. INTRODUCTION

The proposed project focuses on utilizing machine learning techniques to combat the escalating cybersecurity threats associated with malicious URLs. Machine learning, a branch of artificial intelligence, empowers machines to learn from data and previous experiences, identifying patterns and making predictions with minimal human intervention. The project employs supervised machine learning, training machines using a labeled dataset to predict output. The primary objective is classification, a supervised learning technique that categorizes new observations based on training data. By leveraging a multiclass classification technique, the project classifies URLs into four types, namely phishing URLs, defacement URLs, benign URLs, and malware URLs. Three classification algorithms—Random Forest, LightGBM, and XGBoost—are considered, with the most accurate algorithm being selected[7].

Cybersecurity is paramount in safeguarding internet-connected devices and services from malevolent attacks. In today's technology-dependent world, where internet usage is prevalent, cybercrime has seen a significant surge. Malicious URLs serve as a tool for hackers to extract personal information or implant malware on users' devices. The project aims to preemptively detect and predict the harmful nature of URLs to users. When a user clicks on a URL in their browser, the system performs a safety check. If the URL is deemed benign, it can be accessed securely; however, if identified as malicious, an alert message is displayed, informing the user of the URL type. This approach mitigates the risks associated with personal information theft, data loss, and malware downloads, bolstering data privacy, security, and counteracting cyber attacks and crimes.

The project's motivation stems from the imperative need to prevent the proliferation of cyber attacks through social media platforms, protect individuals' personal information, avert data leaks within organizations, and identify malicious URLs based on their appearance. By harnessing lexical features and incorporating them with other pertinent attributes, the project strives to construct a robust model that enhances the detection and classification of malicious URLs.

DOI: 10.48175/IJARSCT-10898

By developing an efficient system capable of accurately identifying and categorizing malicious URLs, the project contributes to fortifying cybersecurity measures, assuring users' data protection, and preserving privacy in the digital realm. The integration of machine learning techniques enables the creation of a proactive defense mechanism against cyber threats, offering a reliable line of defense against phishing attempts, defacement incidents, malware infections, and other forms of cyber attacks. Ultimately, the project aspires to foster a secure and trustworthy online environment, shielding users from the detrimental consequences of cybercrime and reinforcing the resilience of our digital infrastructure.

II. LITERATURE SURVEY

Swati Xess et al. [1] proposed a machine learning approach using the Kaggle dataset. They found that the RandomForest algorithm yielded the best results among the various algorithms tested. This highlights the effectiveness of Random Forest in detecting malicious websites.

Manyumwa et al. [2] compared the performance of ensemble learners including XGBoost, AdaBoost, LightGBM, and CatBoost. They also considered various URL features. The study achieved an overall accuracy above 0.95, indicating the robustness of the ensemble learning approach in detecting malicious URLs.

Rong et al. [3] introduced MalFinder, an ensemble learning-based framework for detecting malicious traffic. They extracted statistical and sequence features from network traffic data and utilized classifiers such as Random Forest, XGBoost, and LightGBM. The proposed system achieved a high F-measure of 0.9346 and 91.04% accuracy in detecting malicious traffic.

Joshi et al. [4] focused on static lexical feature extraction for identifying malicious URLs. They used algorithms including Random Forest, Naive Bayes, SVM, and Logistic Regression as classifiers. The study achieved an accuracy of 0.92, showcasing the effectiveness of the static lexical feature approach in detecting malicious URLs.

Desai et al. [5] developed a Chrome Extension that utilized machine learning algorithms such as kNN, SVM, and Random Forest for identifying phishing websites. After extracting features from the URLs, the Random Forest algorithm achieved the highest accuracy of 0.9611, demonstrating its capability in identifying malicious URLs.

III. PROPOSED SYSTEM

In this section, we present the details of the proposed system, which aims to address the challenge of identifying and categorizing URLs in real-time using machine learning techniques. The system utilizes a Chrome extension that integrates advanced algorithms and features to enhance internet security and prevent cyber-attacks. The key components and functionalities of the proposed system are outlined below.

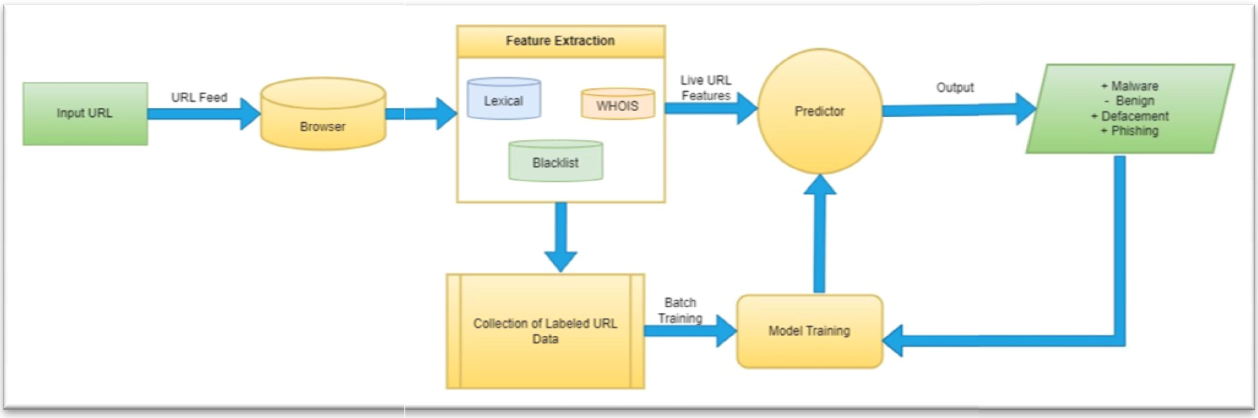

Fig. 1 System Architecture

Data Collection and Preprocessing:

The system starts by collecting a large dataset of URLs from various sources, including malicious, benign, phishing, and spam categories.

The collected URLs undergo a preprocessing stage where relevant features are extracted, such as URL length, domain name, IP address, and content. This step ensures that the data is in a suitable format for training and testing the machine learning models.

Machine Learning Model Selection:

The proposed system employs machine learning algorithms for URL classification, with a focus on ensemble methods like Random Forest, LightGBM, and XGBoost.

These algorithms have been chosen for their ability to handle multi-class classification problems effectively and provide robust predictions.

The system evaluates the performance of each algorithm and selects the best-performing model based on various metrics such as accuracy, precision, recall, and F1-score.

Chrome Extension Development:

To provide real-time analysis and warnings to users, the selected machine learning model is integrated into a Chrome extension.

The extension runs in the background and analyzes the URLs as users browse the internet, leveraging the trained model to classify URLs into different categories.

Upon accessing a potentially harmful URL, the extension triggers a warning message to alert the user and prevent further interaction with the malicious content.

User Interface and Interaction:

The Chrome extension incorporates a user-friendly interface to facilitate seamless user interaction and understanding of the system's functionality.

The interface includes informative elements such as progress indicators, status updates, and clear warnings to enhance user experience and cybersecurity awareness.

Privacy and Security Considerations:

The proposed system prioritizes user privacy and data protection. The collected browsing data is handled securely, adhering to privacy regulations and best practices.

Measures are implemented to ensure that the system does not collect or store any personally identifiable information (PII) during the URL analysis process.

IV. ANALYSIS MODELS: SDLC MODEL TO BE APPLIED

Agile Model : - Why Agile Model?

In our project we are going to use the agile software development model, because in our project as we mentioned we are building a multiclass classification model for identification of harmful URL. So here during use of our webapp if user found some URLs as harmful. These set of URLs will be used as training data which is going to increase efficiency and performance of classificationmodel. So here according the requirement of users we are improving.so agile software development model suits well[6].

Agile model also focuses on direct customer communication instead of making documentation. Due to these environment developer team also get a chance to interact with client so it will beneficial for making a good product and due to these there will be minimum chances of confusion we can also say that the Customer interaction it is an backbone of Agile methodology so due to these all reasons we are using Agile methodology instead of Traditional models.

Steps in the Agile Model:

1. Requirements collection
2. Design

DOI: 10.48175/IJARSCT-10898

3. Iterations of the project
4. Testing the functionality
5. Deployment of the product
6. Feedback from customer

V. CONCLUSION

Machine learning plays vital role in classification and prediction area. It brings revolutionary changes in those area. So in this proposed model, it effectively improved success ratio of systems. Supervised machine learning is used for classifying about type of URL according to the provided dataset. Thus due to this user will get early information regarding unsafe URL. So this will help user to prevent loss of personal information, download of malware, and maintain privacy of user. Thus result in reduce of cyber attacks through unsafe URL. So in this way proposed system is able to satisfy all objectives of project. System is able to identify malicious nature of URL, based in lexical features of URL. Also proposed system is able to prevent cyber attacks and provide security and data privacy.

VI. ACKNOWLEDGMENT

Our heartfelt thanks go to our Principal, Dr. M.A. Venkatesh, for their unwavering support. We are grateful to the Head of Department Prof. R.L. Paikrao, Project Coordinator Dr. R.G. Tambe, and Dr. D.R. Patil for their encouragement. Lastly, we express our deep appreciation and gratitude to our guide, Prof. K.U. Rahane, for their invaluable guidance and constant motivation throughout this project. Additionally, we would like to acknowledge the contributions of our friends, faculty, and staff members. Their support and assistance have been instrumental in our research endeavor. We remain indebted to all who have directly or indirectly aided us in this undertaking.

REFERENCES

[1] Xess, L. S., Khera, M., Prasad, T., Singh, R., & Aiden, M. K. (2022). Malicious Website Detection using Machine Learning. International Journal of Engineering Research & Technology, Published On 2022.

[2] Manyumwa, T., Chapita, P. F., Wu, H., & Ji, S. (2021). Towards Fighting Cybercrime: Malicious URL Attack Type Detection using Multiclass Classification. IEEE, Published On 2021.

[3] Rong, C., Gou, G., Cui, M., Xiong, G., Li, Z., & Guo, L. (2020). MalFinder: An Ensemble Learning-based Framework For Malicious Traffic Detection. IEEE, Published On 2020.

[4] Joshi, A., Lloyd, L., Westin, P., &Seethapathy, S. (2019). Using Lexical Features for Malicious URL Detection - A Machine Learning Approach. ARXIV, Published On 2019.

[5] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017). Malicious Web Content Detection Using Machine Learning. IEEE, Published On 2017.

[6] Murch, R., Milne, J. (2012). System Development and LifeCycle Management (SDLCM) Methodology. United States Nuclear Regulatory Commission, Washington, DC, Vol-3, pages 665.

[7] Mohammed, M., Khan, M. B., Bashier, E. B. M. (2016). Machine Learning: Algorithms and Applications. CRC Press, ISBN: 9781498705394.

DOI: 10.48175/IJARSCT-10898