Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1199

Detection of Malicious Websites using Machine Learning

S Ashok Kumar[1]; Dr D Brindha[2]
URK20CS1010[1]; Assistant Professor[2]
Karunya Institute of Technology and Sciences, Coimbatore

Abstract:- Finding dangerous websites has grown more important as online risks have multiplied in order to protect users' security and privacy. This research uses machine learning techniques to providea new method for spotting dangerous websites. In order to build a strong classifier that can differentiatebetween websites that are harmful and those that arebenign, the suggested approach makes use of a wide range of variables that are taken from user behavior,network traffic, and website content. Analyzing a variety of parameters, including domain age, IP repute, URL structure, HTML content, SSL certificate information, and user interaction patterns,is part of the feature extraction process. These characteristics offer insightful information about the behavior and characteristics of websites, which helps the classifier distinguish between dangerous and legitimate entities.

I. INTRODUCTION

The internet is a fundamental component of contemporary society in the digital age, enabling worldwide communication, trade, and information sharing. However, in addition to all of its advantages, there is a dark side to the internet: bad actors are always looking for ways to use weaknesses for evil. The growth of harmful websites is one of the most pervasive and sneaky types of cybercrime, endangering users' security, privacy, and online well-being.

Phishing websites, virus distribution networks, phony e-commerce portals, and misleading content repositories are just a few examples of the wide range of entities created with harmful intent that fall under the category of dangerous websites. These websites frequently pose as trustworthy organizations in an effort to trick gullible visitors into disclosing private information, downloading malware, or completing fraudulent transactions. For both internet users and cybersecurity experts, identifying and thwarting these attacks is of utmost importance, which calls for the creation of strong and proactive protection systems.

Traditionally, methods for identifying malicious websites have mostly depended on signature-based techniques, in which potentially hazardous content is flagged by matching known malicious patterns againstincoming web traffic. These static and rule-based approaches, while somewhat successful, frequently failto identify new and sophisticated threats that elude signature-based detection systems. Furthermore, because of the internet's dynamic nature and the rogue websites' quick spread, conventional methods are becoming less and less effective in containing new cyber threats.

The application of machine learning techniques to cybersecurity has become a viable paradigm change in response to these difficulties, providing a data-driven and adaptive method for detecting dangerous websites.Machine learning algorithms are an effective tool for evaluating the complex features of websites and differentiating between benign and dangerous entities because of their capacity to learn patterns and draw conclusions from data.

A labeled dataset with instances of both dangerous and benign websites is used to train the detection model. Using supervised learning techniques, a classification model that can reliably identify websites as benign or malicious is constructed. Examples of these techniques include decision trees, random forests, support vector machines, and neural networks. The model is highly accurate and capable of generalizing across various web settings because it has been trained on a substantialamount of labeled data.

A number of essential elements are included in the suggested method, such as feature extraction, dataset preparation, model training, and evaluation. A variety site properties, including HTML content, SSL certificate information, domain age, IP repute, URL structure, and user interaction patterns, are analyzed during the feature extraction process. These qualities offer important insights into the behavior and characteristics of malicious and benign websites, actingas discriminative elements in the process.

Furthermore, sophisticated methods including feature selection, ensemble learning, and model fine-tuning areused to improve the system's functionality and resilience to changing threats. By using these strategies,the classifier's performance is increased, overfitting is decreased, and its capacity to identify malicious patterns that were previously undetected is enhanced.

II. RELATED WORK

Author [1] In the composition" A Phishing Attack What Is It? Determining and Characterizing colorful Phishing AttackTypes" byN. Lord, published in Digital Guardian in 2018, the author is explaining the conception of phishing attacks and agitating colorful types of phishing attacks that individualities and associations may encounter.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1199

Author [2] Machine Literacy algorithms can be trained on labeled datasets of phishing and licit emails to learn patterns and connections between features that distinguish between the two types of emails. The paper may bandy the selection and engineering of features, the choice of machine literacy algorithms, and the evaluation of the performance of the discovery system.

Author [3] The authors likely explore features and characteristics of URLs that can be reflective of vicious intent. These features might include the sphere name, URL structure, presence of suspicious keywords or patterns, hosting information, and literal data on the URL's geste.

Author [4] The author likely explores a range of countermeasures that have been proposed and employed to combat phishing attacks. These countermeasures might include specialized results similar as dispatch filtering, antiphishing toolbars, web runner analysis, and blacklisting of known phishing websites.

Author [5] the authors likely describe the perpetration of associative bracket algorithms, which may include algorithms similar as Apriori, FP- growth, or other association rule mining ways, combined with bracket algorithms like decision trees, neural networks, or Bayesianclassifiers.

III. SYSTEM DESIGN

- Data Collection: Compile information from websitesabout user activities, network traffic, and content. Features including domain age, URL structure, SSL certificates, and HTML content can be extracted. The process of preparing a dataset involves labeling instances of both benign and dangerous websites.
- Model Training: To train a classifier, use supervised learning methods such as neural networks, SVMs, random forests, and decision trees.
- Evaluation: Use metrics such as accuracy, precision, and recall to evaluate the performance of the classifier. Deployment: Use the trained model to automatically identify dangerous URLs in a real-time setting.

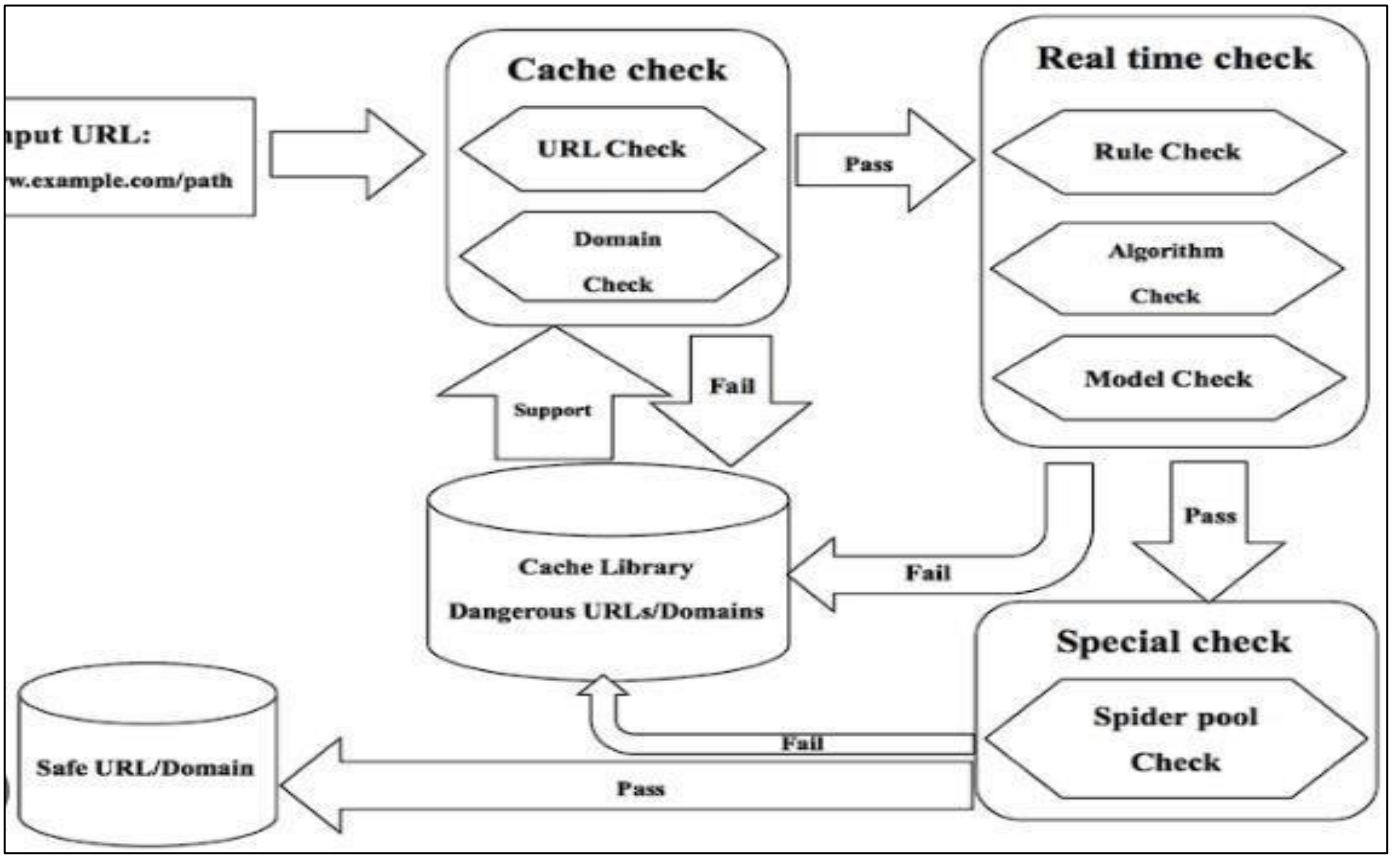

Fig 1: Tool Flowcharts

IV. METHODOLOGY

A. Data Collection:

Compiling a thorough dataset including instances of both dangerous and benign websites is the first stage. To guarantee the efficacy and generalizability of the machine learning model, this dataset ought to be varied and indicative of actual online traffic. The dataset can be gathered from a variety of sources, including web crawlers,threat intelligence feeds, and user-reported data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1199

Nr.	Attribute	Format	Description
1	qty_dot_file	Number of "." signs	Numeric
2	qty_hyphen_file	Number of "-" signs	Numeric
3	qty_underline_file	Number of "_" signs	Numeric
4	qty_slash_file	Number of "/" signs	Numeric
5	qty_questionmark_file	Number of "?" signs	Numeric
6	qty_equal_file	Number of "=" signs	Numeric
7	qty_at_file	Number of "@" signs	Numeric
8	qty_and_file	Number of "&" signs	Numeric
9	qty_exclamation_file	Number of "!" signs	Numeric
10	qty_space_file	Number of " " signs	Numeric
11	qty_tilde_file	Number of "signs	Numeric
12	qty_comma_file	Number of "," signs	Numeric
13	qty_plus_file	Number of "+" signs	Numeric
14	qty_asterisk_file	Number of "*" signs	Numeric

Fig 2: Dataset Atrributes

B. Feature Extraction:

Is the process of taking pertinent features from the attributes of the websites once the dataset has been gathered. Features can be obtained from a number of sources, such as network traffic patterns, SSL certificate information, URL structure, content analysis (such as HTML elements and keywords), domain attributes (such as age, registrar, and popularity), and user behavior (such as clickstream data). The machine learning model is trained using these attributes as input variables.

C. Model Training:

A variety of machine learning techniques can be used to train the detection model using the preprocessed dataset and features that were extracted. For this kind of work, supervised learning techniques like neural networks, decision trees, random forests, and support vector machinesare frequently employed. To reduce prediction errors, the model is trained on the training data using iterative optimization approaches. The dataset is divided into training and testing subsets.

D. Fine-Tuning and Optimization:

Hyperparameter tweaking and optimization strategies can be used to boost the model's efficiency even more. The besthyperparameters that optimize the model's performance metrics can be found using grid search, random search, and Bayesian optimization techniques.

The below URL demonstrates traits that are typically linked to malevolent intent. Its lack of HTTPS encryption raises the possibility of data manipulation or interception. A potential attempt to avoid discovery is suggested by the short domain ageand bad IP reputation, which add to the suspicion. Furthermore,the URL path's inclusion of a login page is a sign of phishing activity, in which users could be duped into disclosing personalinformation. Consequently, based on the combination of these harmful signs, the model properly forecasts this URL as "bad".

Fig 3: Result for Malicious URL

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1199

It uses HTTPS encryption, has an easy-to-understand URL structure, and is registered with a reliable domain registrar. Moreover, the domain age implies that it has been in existencefor a significant duration, adding to its legitimacy. All in all, the characteristics match those generally linked to safe URLs, hence the forecast is favorable.

Neural networks (RNNs), could offer potential improvements in feature representation and detection accuracy. Furthermore, research into collaborative and federated learning approaches could enable the development of more scalable and privacy-preserving detection systems across distributed networks.

V. CONCLUSION

The integration of machine learning techniques holds immense promise for detecting malicious websites, bolstering cybersecurity defenses, and safeguarding users' online experiences. By leveraging comprehensive feature sets and advanced algorithms, machine learning models can effectively discern between benign and malicious entities, offering proactive defense mechanisms against evolving cyber threats. Continuous research and development efforts aimed at enhancing model robustness, scalability, and adaptability will be pivotal in addressing the dynamic nature of malicious activities on the internet. Through collaborative endeavors and innovative approaches, the detection of malicious websites using machine learning stands poised to play a crucial role in ensuring a secure and trustworthyonline environment.

REFERENCES

[1]. N. Lord "A Phishing Attack: What Is It? Determining and Characterizing Various Phishing Attack Types.Digital Guardian, (2018). "What is a phishing attack?": A definition and identification of several forms of phishing attacks.
[2]. The paper "Learning to detect phishing emails" was presented at the 16th International Conference on the World Wide Web in 2007, including papers by N. Sadeh, A. Tomasic, and I. Fette.
[3]. "Learning to detect malicious URLs," ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 9, pp. 30:1-30:24, 2011, J. Ma, S. S. Savag, and G. M. Voelker.
[4]. The article "Phishing counter measures and their effectiveness–literature review" was published in Information Management & Computer Security in 2012.
[5]. The paper "Phishing Detection based Associative Classification" was published in 2014 by N. Abdelhamid, A. Ayesh, and F. Thabtah in Expert Systems with Applications (ESWA), vol. 41, pp. 5948–59.