MALWARE WEBSITE DETECTION USING MACHINE LEARNING

Mr. R. Makendran Assistant Professor, Computer Science and Engineering, Dhirajlal Gandhi College of Technology

Ms. C. Shaheerabanu Student, Computer Science and Engineering, Dhirajlal Gandhi College of Technology

Ms .R. Vinisha Student, Computer Science and Engineering, Dhirajlal Gandhi College of Technology

Ms .P. Sandhiya Student, Computer Science and Engineering, Dhirajlal Gandhi College of Technology

Ms. P. Yogamalya Student, Computer Science and Engineering, Dhirajlal Gandhi College of Technology

---

Abstract - Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. The word "malware" refers to an intent to harm. In order to harm the end user, a malware website spreads malware, infects the victim's system, and steals important information. In the year 2020, the global lockdown saw an increase in and shift toward using internet services as a mode of operation while staying at home. This, in turn, led to an increase in the number of cybercrimes committed by criminals and significant data breaches suffered by businesses. In order to stop these attacks, malware URLs and threat types must be located. Static properties that describe these behaviours can be used to identify the vast majority of malware web pages because they import exploits from distant resources and conceal exploit code. To identify such phishing URLs, a number of models and methods have been proposed in recent years. The previous research is reviewed and a machine learning strategy for the most accurate detection of malware websites using a machine learning model is proposed in this project . In addition, we conduct a reconnaissance on the URL to provide additional information regarding the website's subdomains, directories, and port status.

***Key Words*: Malicious website detection, Feature extraction, Machine Learning**

1.INTRODUCTION

Machine learning is a subset of synthetic intelligence that involves education pc systems to research from facts and improve their performance on a particular task. Machine studying includes various strategies consisting of supervised getting to know, unsupervised gaining knowledge of, and reinforcement getting to know. In supervised learning, the set of rules is trained the usage of labeled information, and the intention is to are expecting an output for brand spanking new, unseen input records correctly. In unsupervised learning, the algorithm is trained using unlabeled statistics, and the aim is to find out styles and shape in the recordsNow a days new Communication Technologies has had a tremendous impact in promotion and business growth spanning across many applications. Unfortunately, the technological advancements come with new sophisticated techniques to attack and scam users. Cybersecurity has been a significant obstacle for computer systems recently because of the rise in Internet usage different malware URLs try to steal user information by releasing various malware programs. Such websites have frequently been identified using signature-based methods, and malware URLs that have been identified have attempted to restrict access by utilizing a variety of security components. There are wide variety of implementations for the attacks such as explicit hacking attempts, drive-by exploits phishing, watering hole, Social engineering, man-in-the middle, SQL injections, loss or theft services etc. Most of these attacking techniques are known through spreading compromised URLs. A URL has two main components, first is the protocol identifier, it indicates what protocol to use and second is the resource name, it specifies the IP address or the domain name where the resource is located. The protocol identifier and the resource name are separated by a colon and two forward slashes.

2 . LITRETURE SURVEY

Phishing is a kind of worldwide spread cybercrime that uses disguised websites to trick users into downloading malware or providing personally sensitive information to attackers. With the rapid development of artificial intelligence, more and more researchers in the cybersecurity field utilize machine learning and deep learning algorithms to classify phishing websites. In order to compare the performances of various machine learning and deep learning methods, several experiments are conducted in this study. According to the experimental results, ensemble machine learning algorithms stand out among other candidates in both detection accuracy and computational consumption. Furthermore, the ensemble architectures still provide impressive capability when the number of features decreases sharply in the dataset. Subsequently, the paper discusses the factors why ensemble machine learning methods are more suitable for the binary phishing classification challenge in up-date training and realtime detecting environment, which reflects the sufficiency of ensemble machine learning methods in anti-phishing techniques.

3. SYSTEM IMPLEMENTATION:

EXISTING SYSTEM:

An existing novel approach for Malware website detection that uses LSTM(long short-termmemory networks)

algorithm .Existing to exploit quantitative data flow properties to extract highly characteristic behavior patterns from collections of known malware.This techniques provide low classification accuracy prediction results. One of the most powerful algorithms in machine learning technology is Random Forest algorithm and it is based on concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy. Random Forests is an Ensemble approach for classification and regression method suitable for handling problems involving grouping of data into classes. It also a supervised machine learning algorithm that built randomly to create a forest. The algorithm was developed by Breiman and Cutler. Classification method use a series of Classification and Regression Tree (CART) to construct the random bootstrap samples of the original data sample.

PROPOSED SYSTEM:

In recent years, a number of models and approaches have been proposed to identify such malware URLs.In this project, a machine learning strategy based on a machine learning model for the most accurate detection of malware websites is reviewed and proposed. Machine Learning Classification algorithm is applied to malware dataset so that we predict the malware.This system is implemented using the following modules.

COLLECTION OF DATASET:

This phase is engineering oriented, which aims to collect most if not all relevant information about the URL. This includes information such as presence of the URLs in a blacklist, the direct features of the URL such as the URL String and information about the host, the content of the web-site such as HTML and JavaScript, popularity information

PRE-PROCESSING OF DATA:

In this phase, the unstructured information about the URL (e.g.,textual description) is appropriately formatted, and converted to a numerical vector so that it can be fed into machine learning algorithms. For example, the numerical information can be used as is, and the Bow is used for representing textual or lexical content. Besides, some data normalization (e.g., Z-score normalization) may often be used to handle the scaling issue.

FEATURE EXTRACTION:

For malicious URL detection, researchers have proposed several types of features that can be used to provide useful information. We categorize these features into: Blacklist Features, URL-based Lexical Features, Host-based features, Content-based Features, and Others. All have their benefits and shortcomings while some are very informative, obtaining these features can be very expensive. Similarly, different features have different preprocessing challenges and security concerns.

LABELLING :

Labelled data is a group of samples that have been tagged with one or more labels. Labelling typically takes a set of unlabeled data and augments each piece of that unlabeled data with meaningful tags that are informative

VALIDATION AND PREDICTION:

Random Forest algorithm avoids overfitting when an enormous number of trees are considered. Random Forest can handle missing values by itself which is the main advantage over other algorithms. The final accuracy of the random forest will be the mean of accuracies generated by all the decision trees

4. SYSTEM ARCHITECTURE

The architecture for malware website detection using machine learning involves collection of the datasets, preprocessing the data and feature extraction process are done for visualizing and prediction by using Random forest algorithm performance by using the datasets. The learning algorithm finds patterns in the training data that map then input data attributes to the target and it outputs an ML model that captures these patterns .Evaluating the models by using the dataset.

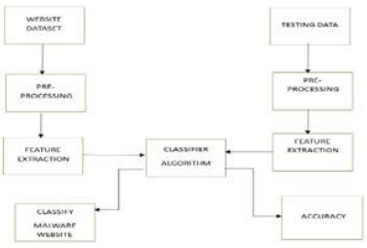

Fig -1: System Architecture

5. CONCLUSION

This project have explored how well to classify phishing URLs from the given set of URLs containing benign and phishing URLs. We have also discussed the randomization of the dataset, feature engineering, feature extraction using lexical analysis hostbased features and statistical analysis. We have also used different classifiers for the comparative study and found that the findings are almost consistent across the different classifiers. We also observed dataset randomization yielded a great optimization and the accuracy of the classifier improved significantly. We have adopted a simple approach to

extract the features from the URLs using simple regular expressions. There could be more features that can be experimented and that might lead to improving further the accuracy of the system. The dataset used in this paper contains the URLs list which may be a little old, hence regular

continuous training along with a new dataset would enhance the model accuracy and performance significantly. In our experiment we have not used the content based features as the main problem with the content-based strategy for detecting phishing URLs is the non-availability of phishing web-sites and the life span of the phishing website is small, and it is difficult to train an ML classifier based on its content-based features.

REFERENCES

[1] Andre Bergholz, Jeong Ho Chang, Gerhard Paaß, Frank Reichartz, and Siehyun Strobel. Improved phishing detection using model-based features. In CEAS, 2008.

[2] Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. The Review of Economic Studies, 78(4):1201–1236, 2021.

[3] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. Spread of (mis) information in social networks. Games and Economic Behavior, 70(2):194–227, 2021

[4] Fabr´ıcio Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virg´ılio Almeida. Characterizing user behavior in online social networks. In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, pages 49–62. ACM, 2021.

[5] G. McGraw and G. Morrisett, “Attacking malicious code: report to the Infosec research council,” IEEE Software, 17(5):33 - 41, Sept./Oct. 2000. [13] S. M. Tabish, M. Z. Shafiq and M. Farooq, “Malware Detection using Statistical Analysis of Byte-Level File Content,” CSI-KDD’09, pp.23-31,

[6] Khammas, Ban Mohammed, et al. "FEATURE SELECTION AND MACHINE LEARNING CLASSIFICATION FOR MALWARE DETECTION." Jurnal Teknologi 77.1 (2015).

[7] Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi. Exposure: Finding malicious domains using passive dns analysis. In NDSS, 2011.

[8] Lu, Yi-Bin, Shu-Chang Din, Chao-Fu Zheng, and Bai Jian Gao. "Using multi-feature and classifier ensembles to improve malware detection." Journal of CCIT 39, no. 2 (2010): 57-72.

[9] Ranveer, S., & Hiray, S. SVM Based Effective Malware Detection System. In: 2015 International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3361-3365.

[10] Ross Anderson and Tyler Moore. The economics of information security. Science, 314(5799):610–613, 2006.

[11] Sulaiman A, Ramamoorthy K, Mukkamala S et al. Disassembled code analyzer for malware. Information Reuse and Integration, Conf, 2005 IEEE International Conference on 2005, 398-403.

[12] Shih-Yao Dai, Sy-Yen Kuo. MAPMon: A Host-Based Malware Detection Tool. The 13th IEEE International Symposium on Pacific Rim Dependable Computing. 2007.