International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Special Issue: | Apr 2023 www.irjet.net p-ISSN: 2395-0072

International Conference on Recent Trends in Engineering & Technology- 2023 (ICRTET-3)
Organised by: VSM College of Engineering, Ramachandrapuram

Malignant URL Detection Using Machine Learning

N.S.C. Mohana Rao[1], G. Vijaya Bharathi[2], B. Karuna[3], Ch. Ephraim Martin[4], J. Nagesh[5], A. Sai Pragathi[6]

[1]Associate Professor,[2-6]Students B.Tech. Computer Science Engineering, V. S. M. College of Engineering,Ramachandrapuram, A.P, India

---

Abstract - Malicious URL detection is an essential component of web security, and traditional rule-based systems have limitations in detecting unknown threats. In recent years, machine learning techniques have been widely used to improve the accuracy of malicious URL detection. In this paper, we propose a machine learning-based approach for detecting malicious URLs. The proposed system uses various features extracted from the URLs, such as the length, entropy, and presence of specific characters, to train a supervised learning model. We compare the performance of various classification algorithms and evaluate the proposed approach on a publicly available dataset. The experimental results show that the proposed system achieves high accuracy in detecting malicious URLs and outperforms state-of-the-art techniques. This approach can be used to develop an effective web security system for protecting users from malicious websites.

Key Words: Malicious, URL, Threats, Websites.

1. INTRODUCTION

Malicious URLs are one of the most significant threats to web security, as they can lead to various forms of cyber attacks such as phishing, malware distribution, and identity theft. Traditional rule-based systems have limitations in detecting unknown threats and often rely on manual updates to their databases. As the number of malicious URLs continues to grow, there is a need for more advanced techniques to detect and prevent them. Machine learning techniques have emerged as a promising solution to this problem.

Machine learning algorithms can learn from large amounts of data and detect patterns that are difficult for humans to identify. In the context of malicious URL detection, machine learning can analyze various features of URLs and classify them as either benign or malicious. These features may include the length of the URL, the presence of certain characters, and the entropy of the URL. By training a machine learning model on a dataset of known malicious and benign URLs, the model can learn to differentiate between them and accurately detect new instances of malicious URLs.

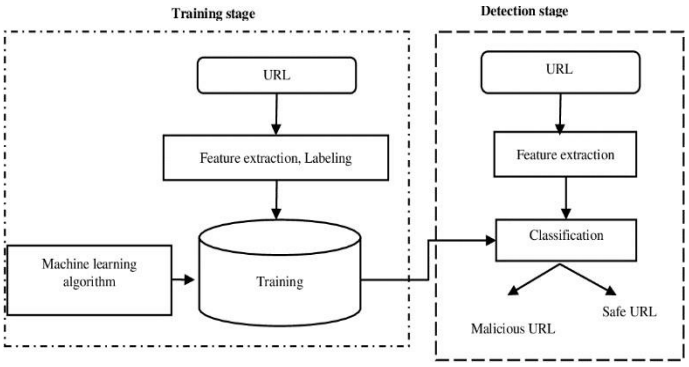

Fig -1: System Architecture

2. IDENTIFY, RESEARCH AND COLLECT IDEA

In [1], M. Zubair Shafiq et al- A Survey of Malicious URL Detection Techniques" in 2018, This survey paper provides an overview of various techniques used for detecting malicious URLs, including rule-based approaches, machine learning, and hybrid approaches.

In [2], Yunsuo Li et al. stated Malicious URL detection based on decision tree - This paper proposes a decision tree-based approach for detecting malicious URLs. It extracts features from URLs and uses decision tree algorithms for classification. The proposed method achieves high accuracy rates in experiments.

In [3], S. Rajendran - This paper proposes a machine learning-based approach for detecting malicious URLs. It uses feature selection techniques to extract relevant features and employs machine learning algorithms such as decision tree and random forest for classification. The proposed approach achieves high accuracy rates in experiments.

In [4], A survey" by Muhammad Salman - This paper provides a comprehensive survey of machine learning-based techniques for detecting malicious URLs. It reviews the challenges and limitations of existing techniques and proposes future research directions.

In [5], A comparative study" by Madiha Khan - This paper presents a comparative study of machine learning-based techniques for detecting malicious URLs. It evaluates the

performance of various machine learning algorithms such as decision tree, and KNN on different datasets.

In [6], A novel feature extraction approach for detecting malicious URLs" by Jian Huang et al. - This paper proposes a novel feature extraction approach for detecting malicious URLs. It uses a graph-based representation of URLs and employs feature extraction techniques such as PageRank and clustering coefficient. The proposed approach achieveshigh accuracy rates in experiments.

3. PROPOSED APPROACH

The proposed malicious URL detection system using machine learning model using machine learning contains two stages, training and detection.

• Training stage: To detect malicious URLs, it is necessary tocollect both malicious URLs and clean URLs. Then, all the malicious and clean URLs are correctly labelled and proceeded to attribute extraction. These attributes will be the best basis for determining which URLs are clean and which are malicious. Details of these attributes will be presented in details in this paper. Finally, this dataset is divided into 2 subsets: training data used for training machine learning algorithms, and testing data used for testing process. If the classification performance of the machine learning model is good (high classification accuracy), the model will be used in the detection phase.

• Detection phase: The detection phase is performed on each input URL. First, the URL will go through attribute extraction process. Next, these attributes are input to the classifier to classify whether the URL is clean or malicious. However, in each attribute group some new attributes and characteristics of the URL to optimize the ability to detect malicious URLs are proposed.

DATA CHARACTERISTICS

In addition to the calories; there are nutrients which needs The URL string contains three different semantic segments, namely: domain name,directory path and file name. The URL and DNS strings consist of numbers, letters and symbols such as “?”, “=”, and “&”. We define the pattern thatcould be used to classify the malicious URLs or normal onesas follows:A URL string is a tuple p = (h, d, f), where h is a URL segment pattern corresponding to the domain name, d= {s1, s2, … sn} is a URL sequential patterns corresponding to the directory path, and f is a URL segment pattern represent the file name. For maliciousURL strings p = (h, d,f) and normal malicious p′ = (h′, d′, f′), if there is a text fragment pattern t′ or other patterns t″ such as URL length is covered by p but not covered by p′, then we view t′ and t″as features which can be used to classify the URL. We seek to automatically find out these features and use them to build an online detection system.

The following are malicious URL examples:
http://www.aaa.com/1.php?Include=http://www.bbb.com/hehe.php
http://www.sqlinsertion.com/adminlogin.php/**/and/**/ 1=1.

4. FUTURE WORK

The proposed work in this paper discusses about the Detection of malicious URL’s by using machine learning algorithms such as Naïve Bayes, Decision tree classifier, SVM, SGD classifier However; the attacking technique’s evaluated day by day so it is need to improve the detection techniques too, to stop these type of attacks.

5. CONCLUSION

In conclusion, malicious URL detection using machine learning has been an active area of research for many years, and significant progress has been made in developing effective approaches for detecting malicious URLs. Various machine learning techniques, such as SVM, decision trees, KNN, and deep learning, have been used for classification, and feature selection techniques, graph-based representations of URLs, and word embeddings have been used for feature extraction.

The results of experiments show that these approaches can achieve high accuracy rates in detecting malicious URLs, with some achieving accuracy rates above 99%. However, there are still challenges in this field, such as the presence of polymorphic malware and the need for timely updates to the training datasets.

ACKNOWLEDGEMENT

We are appreciative to our Division of Computer Science Engineering for their help and giving us a chance to make things simpler. While looking about this point we found out about different significant and fascinating realities. The tools and the web services provided are of much help.

REFERENCES

[1] D. Sahoo, C. Liu, S.C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”. CoRR, abs/1701.07179, 2017.

[2] M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp.2091–2121, 2013.

[3] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drive by download attacks and malicious java script code," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 281–290.

[4] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.

[5] Internet Security Threat Report (ISTR) 2019–Symantec. https://www.symantec.com/content/dam/symantec/docs /reports/istr-24-2019-en.pdf [Last accessed 10/2019].

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang,"An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[7] C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 91–96.

[8] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.

[9] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious urls: an application of large-scale online learning," in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, pp. 681–688.

[10] B. Eshete, A. Villafiorita, and K. Weldemariam, "Binspect: Holistic analysis and detection of malicious web pages," in Security and Privacy in Communication Networks. Springer, 2013, pp. 149–166.

[11] S. Purkait, "Phishing counter measures and their effectiveness– literature review," Information Management & Computer Security, vol. 20, no. 5,pp. 382–420, 2012.

[12] Y. Tao, "Suspicious url and device detection by log mining," Ph.D.dissertation, Applied Sciences: School of Computing Science, 2014.

[13] G. Canfora, E. Medvet, F. Mercaldo, and C. A. Visaggio, "Detection of malicious web pages using system calls sequences," in Availability,Reliability, and Security in Information Systems. Springer, 2014, pp.226–238.

[14] Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32,(2001).

[15] Thomas G. Dietterich. Ensemble Methods in Machine Learning.International Workshop on Multiple Classifier Systems, pp 1-15, Cagliari, Italy, 2000.

[16] Developer Information. https://www.phishtank.com/developer_info.php.[Last accessed 11/2019].

[17] URLhaus Database Dump. https://urlhaus.abuse.ch/downloads/csv/.[Ngày truy nhập 11/2019].

[18] Dataset URL. http://downloads.majestic.com/majestic_million.csv.[Last accessed 10/2019].

[19] Malicious_n_Non-MaliciousURL. https://www.kaggle.com/antonyj453/urldataset#data.csv . [Last accessed 11/2019].

[20] chrome.zip. https://drive.google.com/file/d/13G_Ndr4hMFx_qWyTEjHuOyJmHFWD0Gud/view?fbclid=IwAR0SLVCrvjHHGmoHZH97nXN3Bm-DMY7jG4SOsKZYLAZjTFgeoJADfli64-g. [Lastaccessed 12/2019].

BIOGRAPHIES

Team Guide, N.S.C. Mohana Rao is working as Associate Professor – CSE Department, VSM College of Engineering, Ramachandrapuram.

Team Leader, G. Vijaya Bharathi is B.Tech student, CSE Department, VSM College of Engineering, Ramachandrapuram.

Team Member, B. Karunais B.Tech student, CSE Department, VSM Collegeof Engineering, Ramachandrapuram.

Team Member, Ch. Ephraim Martin is B.Tech student, CSE Department, VSM College of Engineering, Ramachandrapuram.

Team Member, J. Nageshis B.Tech student, CSE Department, VSM Collegeof Engineering, Ramachandrapuram.

Team Member, A. Sai Pragathi is B.Tech student,CSE Department,VSM College of Engineering, Ramachandrapuram.