iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41084

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Implementation Paper on Detection of Malicious URLs Using Machine Learning Techniques

Miss. Mayuri Arvind Pohane[1], Dr. A.A. Bardekar[2]
[1]*PG Scholar,* [2]*Professor, Computer Science & Engineering, Sipna College Of Engineering and Technology, Amravati, Maharashtra*

Abstract: ***Detecting and preventing the user from the malicious site attacks are significant tasks. A huge number of attacks have been observed in last few years. Malicious attack detection and prevention system plays an immense role against these attacks by protecting the system's critical information. The internet security software and fire walls are not enough to provide full protection to the system. Hence efficient detection systems are essential for web security. These existing methods have some drawbacks results into numbers of victims to increase. Hence we developed a system which helps the user to identify whether the website is malicious or not. Our system identifies whether the site is malicious or not through URL.***
Keywords: Malicious URLs, Classifier, Feature Extraction, ID3 Algorithm

I. INTRODUCTION

The data set used in this research consisted of known malicious and known benign URLs. For malicious dataset, we obtained the data from Squid blacklist, a web service which provides lists of malicious URLS. Benign dataset was obtained from Alexa, which is a web service that ranks web sites based on traffic generated. For example, web sites such Google and Facebook will be ranked higher in their list of websites compared those that are less frequented. Higher traffic generated web sites are less likely to be malicious because such sites are well preserved due to their fame among internet. The collected data was grouped into training with 460 instances and 534 attributes and testing data set. Training set is defined as a set of data used to discover potentially predictive relationships. The training set is used at the initial stage of the proposal to determine patterns or similarities between the different set of data obtained. The testing set is used to verify the set of patterns or similarities that was exposed during the training stage. The output of testing includes 220 malicious URL and 240 benign URL.

II. LITERATURE SURVEY

1) They will conduct an experimental user study to evaluate the effectiveness of image comparison and display customization as techniques for users to identify remote servers. We are currently designing a between-subjects study to compare our prototype to other techniques. In the study, participants will be asked to create an account on a remote server and to login. We will periodically send the users email that asks them to login to the website in order manage their funds (which is their payment for participation in the study). Occasionally, users will be sent to a website that spoofs the content of the site as well as the security indicators (such as their trusted window). Participants will be divided into three groups, one using Dynamic Security Skins, one using a shared secret scheme and another using only a standard SSL equipped browser (the control condition). Effectiveness of the prototype will be measured by the performance and error rate in account creation and login tasks, the ability for users to authenticate legitimate servers, the rate of detecting spoof attempts and user satisfaction. Additionally, we plan to release the application to the public for widespread testing

2) This survey provides a system review of extensive research on phishing techniques and countermeasures. Previous surveys and taxonomies either concentrate on one specific aspect of phishing such as anti-phishing tools (Abbasi et al. 2010; Zhang et al. 2011a), or fail to provide an integrated overview of research approaches to various phishing techniques (Huajun et al. 2009; Wetzel 2005; Ollmann 2007a); The taxonomy proposed in this research is multi-dimensional, which distinguishes itself from the previous ones that are focused on a single dimension. In addition, the phishing environment covered in existing taxonomies is limited to traditional channels such as e-mails and spoofed websites

3) They have described an approach for classifying URLs automatically as either malicious or benign based on supervised learning across both lexical and host-based features. We argue that this approach is complementary to both blacklisting — which cannot predict the status of previously unseen URLs — and systems based on evaluating site content and behavior — which require visiting potentially dangerous sites. Further, we show that with appropriate classifiers it is feasible to automatically sift through comprehensive feature sets (i.e., without requiring domain expertise) and identify the most predictive features for classification. An open issue is how to scale our approach to handle millions of URLs whose features evolve over time. We address the issue in subsequent work by using online learning algorithms.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

4) They have developed a system for detection of malicious websites through URL which based on an automated classifier. The classifier is trained with the dataset of legitimate and malicious websites. The trained classifier is for the detection of any URL. Further, the accuracy of the system increases as the classifier is trained with more data set.

III. SYSTEM DIAGRAMS

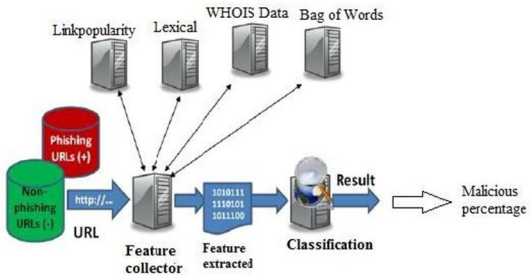


Fig : Learning to detect malicious web sites from suspicious URLs.

IV. PROPOSED WORK

Proposed system is a framework that detects the malicious URL. It takes URL as an input and gives the malicious percentage of URL. Collection of dataset is the very task of implementation. System uses different databases is blacklisting, Bag of Words.

A. Basic Steps of Implementation

1) *Stage 1:* consist of training data collection,
2) *Stage 2:* supervised learning with the training data,
3) *Stage 3:* Malicious URL detection and attack type Identification.

These stages can operate consecutively as in batched learning, or in an interleaving manner: additional data is collected to incrementally train the classification model while the model is used in detection and identification. There are different feature extractions techniques are used like lexical, link popularity, WHOIS, bag of words. The URL goes through all these features and every feature is extracted by machine learning Mechanism. N-gram classifier is used to classify the extracted features. System gives the result in percentage that how many percent the URL is malicious

V. ALGORITHM USED

A. N-GRAM Algorithm

1) *Step 1:* Given a web P, extract its URL identity and generate features.
2) *Step 2:* Classify P by classifier and return result (+1, or 0). //0: legitimate, 1: malicious
3) *Step 3:* Use web feature such as anchor, lexical, WHOIS features extractions If result= 1, output the malicious label, if result=0, go to Step 4.
4) *Step 4:* If P has not a text input, output the malicious label (1).
5) *Step 5:* Compares the criteria with value 1 and 0.
6) *Step 6:* Result, the safe or unsafe percentage of the input.

An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n-1) order. N-gram models are now widely used in probability, communication theory, computational linguistics, computational biology and data compression. We have used this model in proposed system for computational linguistic purpose. Two benefits of n-gram algorithm are simplicity and scalability. An n-gram is a contiguous sequence of n items from given sample of text. The items can be phonemes, syllables, letters, words or base pairs according to the application. When the items are words, n-grams may also be called shingles. Because of its simplicity the n-gram algorithm has been chosen for application of this model.

VI. CONCLUSION

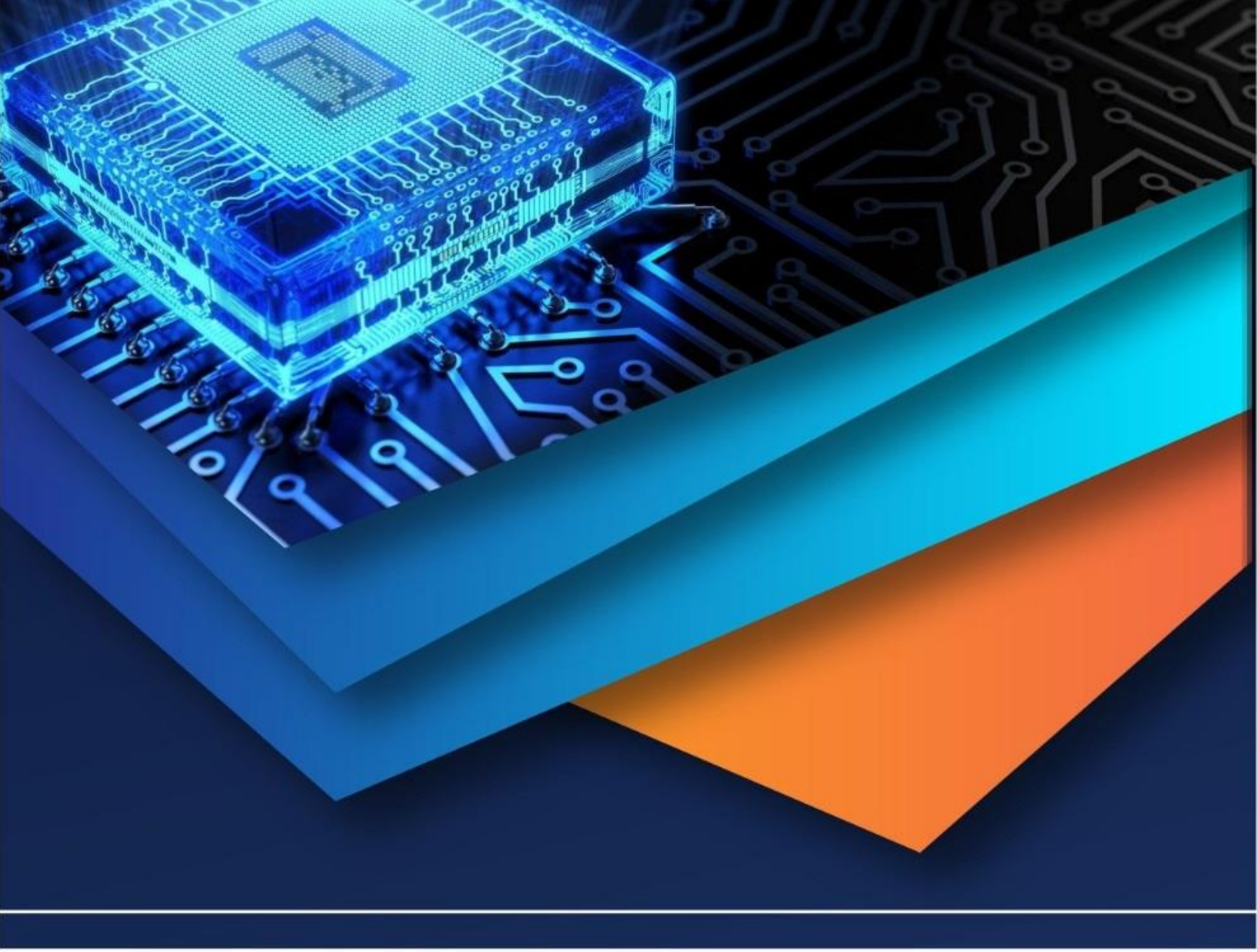
Thus we conclude a system to improve the efficiency of existing system combination of multiple features has been used such as link popularity, lexical, WHOIS, bag of words. System provides the percentage of malicious and benign Links. We will additionally work on providing solution to the problem. The trained classifier is for the detection of any URL. Further, the accuracy of the system

VII. ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my **PROF. A.A. Bardekar** who has in the literal sense, guided and supervised me. I am indebted with a deep sense of gratitude for the sconstant inspiration and valuable guidance throughout the work.

REFERENCES

[1] S. Dhamija, R., and Tygar, J., "The battle against phishing: Dynamic security skins", In Proc. ACM Symposium on Usable Security and Privacy (SOUPS 2005), pp.77–88,.
[2] Chandrasekaran, M., Narayanan, K., and Upadhyaya, S., "Phishing email detection based on structural properties", Proceedings of the NYS Cyber Security Conference, 2006.
[3] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs", In: Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009, pp. 1245-1254.
[4] Yogesh Dubey , Palghar Pranil Chaudhari Student ,Tina D'abreo Lecturer Efficient Detection of Legitimate and Malicious URLs using ID3 Algorithm
[5] Ma, Justin, et al. "Beyond Blacklists: learning to detect malicious websites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009
[6] Jin-Lee Lee,Doung-Hyun Kim,Chang-Hoon Lee. "Heuristic-based Approach for Phishing Site Detection Using URL Features" Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015.
[7] Sana Ansari and Jayant Gadge. "Architecture for Checking Trustworthiness of Websites "International journal of computer application, Volume 44, April 2012
[8] Mustafa Aydin and Nazife Baykal "Feature Extraction and Classification Phishing Websites Based on URL" Cyber Defence and Security Laboratory of METUCOMODO, IEEE CNS 2015.
[9] Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.
[10] Canali, Davide, et al. "Prophiler: a fast filter for the large-scale detection of malicious web pages." Proceedings of the 20th international conference on World wide web. ACM, 2011.
[11] Sumalatha Ramachandran, Sujaya Paulraj, Sharon Joseph and Vetriselvi Ramaraj, "Enhanced Trustworthy and High-Quality Information Retrieval System for Web Search Engines", IJCSI International Journal of Computer Science Issues, Vol. 5, October 2009, pp38- 42.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY