SCD_Report.pdf

by

Submission date: 01-Feb-2022 12:42AM (UTC+0500)
Submission ID: 1752180691
File name: SCD_Report.pdf (301.35K)
Word count: 2606
Character count: 14540

XSecure : Machine Learning Based Phishing Malware Detection

Rehan Mumtaz, Sufiyan Irfan, Kabeer Ahmed
NED University of Engineering and Technology

Abstract— Phishing is a type of online social engineering attack that aims to steal a person's digital identity by impersonating a legitimate entity [1]. The hacker transmits a signal or a message through an email or a link embedded in that email , chat, social media post, or other channel that contains a URL to a malicious website gathering private and sensitive information of the users. We are concentrating and centralizing on developing a system for URL analysis and its classification in order to detect phishing attacks and cyber crimes.This document highlights the requirements and challenges for establishing Malicious URL Detection as a service for real-world cybersecurity applications. So malware website detection is the need of today's internet platform and applications of malware detection should be user-friendly and easily accessible to all users.

Keywords— malicious URL detection; feature classification ,feature extraction; logistic regression, randomforest,

I. INTRODUCTION

Hackers frequently use spam and phishing to trick users into clicking malicious URLs, after which Trojans are installed on the victims' computers or sensitive information is leaked. Malicious URL detection technology can assist users in identifying malicious URLs and preventing them from being attacked by malicious URLs. Blacklist-based methods were used previously to detect malicious URLs. This method has a few distinct advantages. It is fast, has a low false-positive rate, and is simple to implement.

To identify malicious URLs, researchers used a machine learning technique. These methods frequently require manual extraction of the features, and to avoid detection attackers can redesign the features. In today's complex network environment, developing a more effective malicious URL detection model becomes a research priority.

This document proposes a malicious URL detection model based on a logistic regression model using machine learning techniques. This machine Learning security solution, supports an always on detection system, Django framework is used for the backend and simple HTML, CSS used for frontend and designing the pages.

In this study, our innovations and contributions are as follows:

(1) This document demonstrates the use of malicious URL detection model based on a Logistic Regression Model.

(2) In the stage of feature extraction and representation, the features are extracted from the URL sequence [4]. The features classified and extracted are passed into a

vector, and the vector is processed directly by machine learning to learn the classification model. Extracting features like "@" in the URL can redirect you to some other malicious website so if there is an "@" in the URL link the trained model will detect the website as malicious.

(3) To prove the feasibility of the model proposed in this paper, we did a lot of comparative experiments. As for the embedding method, we conduct six contrast experiments such as status alexa index, status abnormal url, status domain age. status redirects, status prefix suffix, status '@' : not applicable. If either of these has a false status then it means the website is not safe for browsing.

II. RELATED WORK

A. Signature based Malicious URL

Detection Studies on malicious URL detection using signature [6] sets have been researched and applied for a long time. The majority of these studies frequently make use of lists of known malicious URLs. A database query is run whenever a new URL is accessed [14,15]. If it is present in the list of blacklisted URLs, then it is malicious and an alert or warning will be generated; otherwise, URLs are considered safe. The main disadvantage of this [8,9] approach is that it will be extremely difficult to detect new malicious URLs that are not included in the provided list.

B. Machine Learning based Malicious URL Detection

Three kinds of machine learning algorithms that can be used to detect malicious URLs: supervised learning, unsupervised learning, and semi supervised learning [13,16]. And the methods of detection are based on URL behaviours. [2,4,5] investigates a number of malicious URL systems based on machine learning algorithms. SVM, Logistic Regression, Naive Bayes, Decision Trees, Randomforest, and other machine learning algorithms are examples. The Logistic Regression Model is used in this paper. The experimental results will show the accuracy of this algorithm with different parameter setups.

There are two categories in which URL characteristics and behaviors can be divided static and dynamic. Authors presented methods for analyzing and extracting static behavior of URLs in their studies [12, 13, 19], including Lexical, Content, Host, and Popularity-based. Online Learning algorithms and SVM were used in these studies as machine learning algorithms. [21, 22] present malicious URL detection using dynamic URL actions. Character and semantic groups are investigated [8], as well as the Abnormal group in websites and the Host-based group; the Correlated group [24].

C. Malicious URL Detection Tools

- URL Void: Program that employs multiple engines and domain range of other testing services.

- UnMask Parasites: Unmask Parasites is a URL testing tool that downloads provided links and parses HTML codes, particularly external links, iframes, and JavaScript. This tool has the advantage of detecting iframes quickly and accurately.

- Comodo Site Inspector: This is a tool for detecting malware and security holes. This allows users to check URLs or webmasters to set up daily checks by downloading all of the specified sites. and run them in a browser sandbox environment

- Other tools: UnShorten.it, Norton Safe Web, Sucuri, VirusTotal, SiteAdvisor, Browser Defender, Online Link and Google Safe Browsing Diagnostic. The majority of current malicious URL detection tools [16] are signature-based URL detection systems, according to the analysis and evaluation of malicious URL detection tools presented above.

blacklists. Google SafeBrowsing, Norton SafeWeb, and MyWOT are some URL Void examples. The Void URL tool has the advantage of being

III. METHODOLOGY

A. The Model

Figure 1 depicts the proposed machine learning-based malicious URL detection system. This model has two stages: training and detection.

I. Extraction Phase & Training stage: To detect malicious URLs, both malicious and legitimate URLs must be collected. From kaggle, we obtained a large dataset constituting around more than 4,00,000+ urls. We gathered around this large dataset of training the model as powerful so that it can give correct results as possible.

Now here comes the feature Extraction step, where the Url is further broken down into the matching 11 listed below protocols and then maintain a vector array to which the match is there If it matches the tokens , it will list the certain token as -1 or otherwise as 1. After those 11 protocols columns there is an other column of Result which tell -1 for Malware and 1 for Legitimitate Url

These comprised vectors are dumped in CSV format and then it get feeded to the training phase as an input for algorithm .The main comprised tokens are

1. URL Length: Lengthy Url makes it ambiguous for the user to check , the user only checks for the main domain and let it go the other path

2.Contain @ Symbols : Most of the time, it happens like '@' is there in the url and no one notices as the browser normally executes the link after '@' symbol

3. Two and Three-Tier Subdomains : Normally Url do not have more than 3 domains under the main name embedded in the URL

4. Suffix & Prefix Separated Domain : Legitimate URLs rarely contain the dash (-) sign. To make a phishing website look like a legitimate website, prefixes or suffixes separated by (-) might be added**.**

5. IP address in the URL : To deceive users and steal important information, an IP address can be substituted for the domain name.

6. '//' makes Redirection : It is possible to determine whether there is any unneeded redirection to other websites by looking at the placement of the '//' in the URL.

7. URL Shortening : Shortening of Url is a technique used on the "World Wide Web" that allows a URL to be significantly reduced in length while still directing to the desired web page.

8. Https Webapp : This token can be used by attackers to deceive users.

9. Traffic Control Web : The popularity or ranking of a website influences whether it is a phishing website or not.

10. Domain Life Period : Phishing websites usually only exist for a short time. Some respectable websites are active for at least 6 months, while the majority are active for at least a year.

11. Database Records : Check whether the URL is listed in the phishing URL lists supplied by PhishTank and others.

II. Detection phase: Following feature extraction, the algorithm receives the processed data set as input, and their results are compared.

After Successful training of the system, the later part which has been left off that is 10% is used to test the accuracy and precision of the model's algorithm

The input to the machine learning algorithms comes from CSV files for both training and testing, with training accounting for 90% of the total URLs in the data set and testing accounting for 10%.

The algorithms used are random forest and logistic regression.

Each input URL is subjected to the detection phase. The URL will first be subjected to the attribute extraction process. Following that, these attributes are fed into the classifier, which determines whether the URL is Legitimate or malicious.

IV. RESULT

We have tested this application for suspected urls published on phishtank and tailed results with Google's Safe Browsing technology.

1.Alexa Ranking :
With Alexa ranking we are getting the ranking of that specific URL whenever it is searched:
Lower the rank more the probability of being legitimate.
Higher the rank the more the probability of being Malicious.

2.Domain registration:
We have marked URLs with registration time of more than 6 months as safe. URLs must qualify for at least one category to be safe either it should have domain registration with more than 6 months registration or it should have lower alexa ranking.

3. Dataset of Malicious_n_Non-Malicious URL [17]: data source with 400,000+ labeled URL. This database contains 83 percent of all URLs that are safe, while the remaining 17 percent are malicious.
The above-mentioned dataset of both safe and malicious URLs is divided into two subsets. Approximately 80% of the dataset is used for training, while the remaining 20% is used for testing.For our machine learning approach, we used a number of classifiers, trained the model on those classifiers and obtained different results [23]. Specifically we have used Logistic Regression (LR) and RandomForest classifier. We then calculated different metrics based on these classifiers like accuracy, recall, precision and f1-score. These metrics have been calculated by using the following equations.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

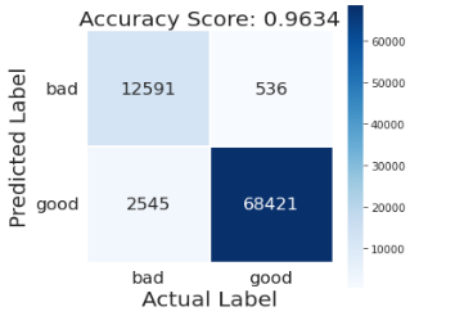
$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

$$F1 - Score = \frac{2\,(Precision \times Recall)}{(Precision + Recall)}$$

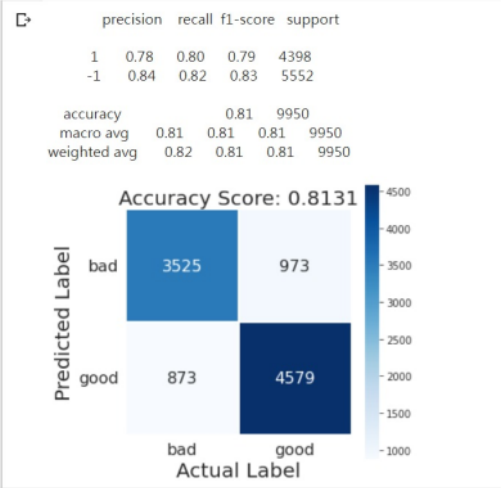
The results from these classifiers are shown below.

1) Logistic Regression:

	precision	recall	f1-score	support
bad	0.96	0.83	0.89	15136
good	0.96	0.99	0.98	68957
accuracy			0.96	84093
macro avg	0.96	0.91	0.93	84093
weighted avg	0.96	0.96	0.96	84093



2) RandomForest Classifier:



As can be seen from the above two tables, the logistic regression model gave the best results. The logistic regression model produced highest precision score of 96%, highest recall with score of 91% and the highest f1-score of 93.5%

V. CONCLUSION

Many cybersecurity applications rely on malicious URL detection, and machine learning algorithms are definitely a promising future. The needs and obstacles for establishing Malicious URL Detection as a service for real-world cybersecurity applications have been highlighted in this document.

Our solution will detect malicious links and its origin signature (first uploaded person-profile URL, name, email, number etc.) on a real time basis through Chrome Extension and provide advisory reports to the public corresponding agencies about those links source credibility. We determine whether a url is malicious or legitimate using Machine Learning.

We might not be perfect but we have a sharp edge over every other similar application due to highly accurate results, Multi-platform support and various other functionalities. This application has the ability that it can be integrated with the other modules which will be created against the cyber attacks in the upcoming cyber era.

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] 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.

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

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

[6] A. Blum, B. Wardman, T. Solorio, and G. Warner. Lexical feature based phishing url detection using online learning. In AISec, pages 54--60, 2010.

[7] Li Xu, Zhenxin Zhan, Shouhuai Xu, and Keying Ye. 2013. Cross-layer detection of malicious websites. In Proceedings of the third ACM conference on Data and application security and privacy. ACM.

[8] Sandeep Yadav, Ashwath Kumar Krishna Reddy, AL Reddy, and Supranamaya Ranjan. 2010. Detecting algorithmically

[9] Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In

[10] 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.

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

[12] 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].

[13] Wei Zhang, REN Huan, and Qingshan Jiang. 2016. Application of Feature Engineering for Phishing Detection. IEICE TRANSACTIONS on Information and Systems (2016).

generated malicious domain names. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM.

[14] Wei Wang and Kenneth Shirley. 2015. Breaking Bad: Detecting malicious domains using word segmentation. arXiv preprint arXiv:1506.04111 (2015).

[15] Yao Wang, Wan-dong Cai, and Peng-cheng Wei. 2016. A deep learning approach for detecting malicious JavaScript code. Security and Communication Networks (2016).

Proceedings of the 25th international conference on Machine learning. ACM.

[16] Weibo Chu, Bin B Zhu, Feng Xue, Xiaohong Guan, and Zhongmin Cai. 2013.

Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs. In Communications (ICC), 2013 IEEE International Conference on. IEEE.

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

[18] Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin SC Yong, and Wei King Tiong. 2019. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences 484 (2019), 153–166

[19] Gerardo Canfora and Corrado Aaron Visaggio. 2016. A set of features to detect web security threats. Journal of Computer Virology and Hacking Techniques (2016).

[20] Eduardo Benavides, Walter Fuertes, Sandra Sanchez, and Manuel Sanchez. 2019. Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review. In Developments and Advances in Defense and Security. Springer, 51–64.

[21] Sreyasee Das Bhattacharjee, Ashit Talukder, Ehab Al-Shaer, and Pratik Doshi. 2017. Prioritized active learning for malicious URL detection using weighted text-based features. In Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on. IEEE.

[22] Yazan Alshboul, Raj Nepali, and Yong Wang. 2015. Detecting malicious short URLs on Twitter. (2015).

[23] Betul Altay, Tansel Dokeroglu, and Ahmet Cosar. 2018. Context-sensitive and keyword density-based supervised machine learning techniques for malicious webpage detection. Soft Computing (2018).

[24] Ankesh Anand, Kshitij Gorde, Joel Ruben Antony Moniz, Noseong Park, Tanmoy Chakraborty, and Bei-Tseng Chu. 2018. Phishing URL detection with oversampling based on text generative adversarial networks. In 2018 IEEE International Conference on Big Data (Big Data). IEEE, 1168–1177.

[25] Farhan Douksieh Abdi and Lian Wenjuan. 2017. Malicious URL Detection using Convolutional Neural Network. Journal International Journal of Computer Science, Engineering and Information Technology (2017).

SCD_Report.pdf

ORIGINALITY REPORT

31%	26%	26%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	thesai.org Internet Source	15%
2	www.hindawi.com Internet Source	6%
3	www.jetir.org Internet Source	3%
4	Cho Do Xuan, Hoa Dinh, Tisenko Victor. "Malicious URL Detection based on Machine Learning", International Journal of Advanced Computer Science and Applications, 2020 Publication	2%
5	Zhiqiang Wang, Xiaorui Ren, Shuhao Li, Bingyan Wang, Jianyi Zhang, Tao Yang. "A Malicious URL Detection Model Based on Convolutional Neural Network", Security and Communication Networks, 2021 Publication	1%
6	Rakesh Verma, Avisha Das. "What's in a URL", Proceedings of the 3rd ACM on International	1%

	Workshop on Security And PrivacyAnalytics - IWSPA '17, 2017 Publication	
7	export.arxiv.org Internet Source	1 %
8	Submitted to Kennesaw State University Student Paper	1 %
9	Submitted to Higher Education Commission Pakistan Student Paper	1 %
10	ijarcce.com Internet Source	1 %
11	Nabeel Al-Milli, Bassam H. Hammo. "A Convolutional Neural Network Model to Detect Illegitimate URLs", 2020 11th International Conference on Information and Communication Systems (ICICS), 2020 Publication	<1 %
12	pure.tue.nl Internet Source	<1 %
13	ijirset.com Internet Source	<1 %
14	"New Trends in Computational Vision and Bio-inspired Computing", Springer Science and Business Media LLC, 2020 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

SCD_Report.pdf

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8