iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VII **Month of publication:** July 2019

DOI: http://doi.org/10.22214/ijraset.2019.7214

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: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

Suspicious URL Detection using Dynamic Learning Model with Machine Learning

Sanjana Tiwari[1], Ashok Kumar Behera[2], Monika Verma[3]
[1]*Computer Science Department, BIT Durg, Chhattisgarh, India*
[2, 3]*Bhilai Institute Of Technology, Durg*

Abstract: Malicious websites that bargain defenseless PCs are an ever-present risk on the web. They are the foundation of Internet criminal exercises and a generally perceived risk to the security of the web and a developing security worry on the Internet because of their popularity and their potential genuine effects. They can 'shroud' the content of the web pages, i.e., serving distinctive content to various customers.
The perils of these sites have made an interest for safeguards that shield end-users from visiting them. At the point when a powerless client utilizes a web program to surf a URL, a vindictive server can send to the web program a web page with malignant code to exploit and bargain the customer side framework.

I. INTRODUCTION

The web has turned into a key worldwide stage and a basic part of the general public that pastes together day by day correspondence, sharing, exchanging, joint effort, and administration conveyance. Countless associations overall depend on the web for their day by day activities, either totally or just too some degree.

These days, the trust in an association vigorously relies upon the nature of its web nearness, which must pass on a feeling of trust and steadfastness to its users over the time. As of late, client user has turned into the fundamental target for attacks, as the enemy trust that the end client is the weakest connection in the security chain. A.R Nagaonkar et al 2016 [16]**,** "Finding the malicious URL using search engine mechanism".

Author uses different types of method for finding the malicious urls they uses SEO method, link based method, DNS query methods, domain registration methods. Basically author combines lexical and host based features to obtain the accuracy.N. Provos et al 2006,[17]

"The Ghost in the Browser," in this paper Author gives the current condition of malware in the internet. The four keypoints outsiders gadgets, promoting, web server security, client contributed substance.

All this fetures get combined and used for internet browser services. This paper is only HTML based and JAVA script Based.P. Mavrommatis et al,[18]"All Your IFrame point to us". In this paper only HTML based feature based is used like IFrame.When landing site wants to interact with drive-by-download victim.Client visits the landing site Redirects to get exploit download the malware executable.J Nazario et al[19]"A Virtual Client Honey Pot" In this paper the author examines about a virtual pot. This paper is only HTML based and JAVA script Based.

II. METHODOLOGY

In this segment we defines how our proposed system works As Figure 2.1shows. To demonstrate this approach, we will built a URL classification system that uses a ICML-2009 datasets.it contains approximately 2.4 million url's and 3.2 million url's features.

Using this data,we will extract the features like lexical features(or we can say the printed features of url).lexical features incorporates length of hostnames and url.

Host based Features tell "where" is the malicious url destination "their identity" survey by and "How" they are govern. web content based features is the combine feature of host based and lexical based feature.the web content based feature uproot by seize the html page of requested site.it includes HTML count, Hyper link count, Iframe count, Suspicious javascript function count. For training and classification of datasets we use three different classifiers: Linear SVM Classifier, K Nearest Neighbors Classifier, Random Forest Classifier.

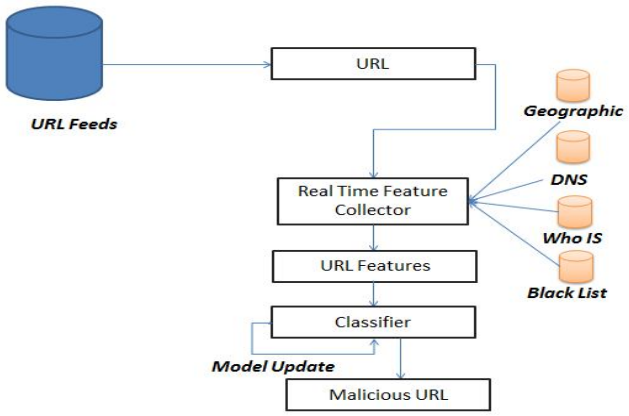


Fig2.1: Proposed system Architecture

III. RESULT

Run Algorithm to trainee datasets by using svm algorithm. As Figure 3.1shows After that svm algorithm run and provides iterations.

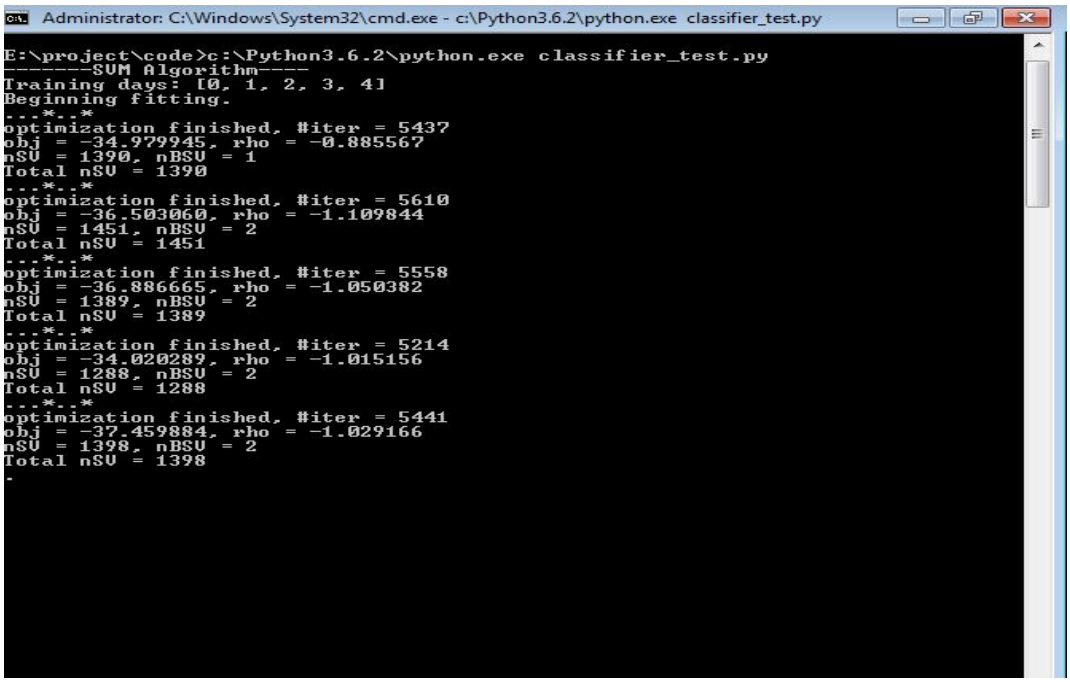


Fig3.1: Training datasets by svm algorithm

Now after completing iterations we achieve accuracy about 97% As Figure 3.2shows by using SVM optimization. By using KNN algorithm we achieve 92% As shown in Figure 3.3, and by using Random forest algorithm we achieve 95% As shown in Figure 3.3.

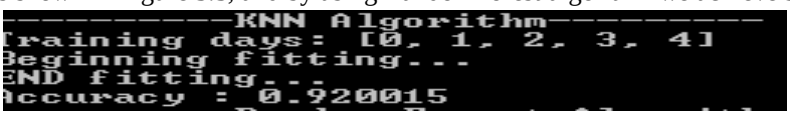


Fig3.2: Training and testing using KNN algorithm

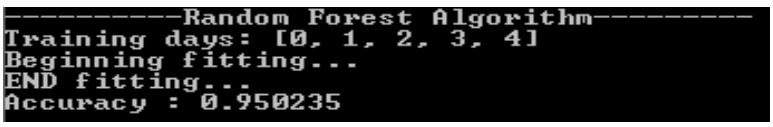


Fig3.3: Training and testing using Random forest algorithm

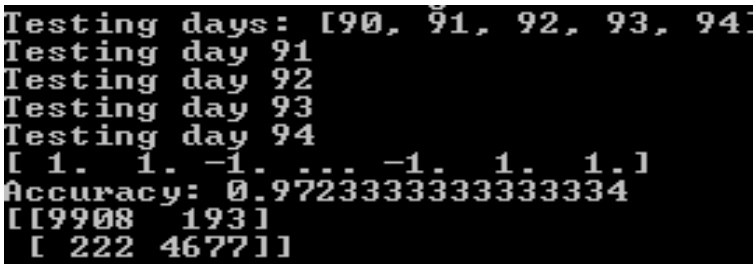
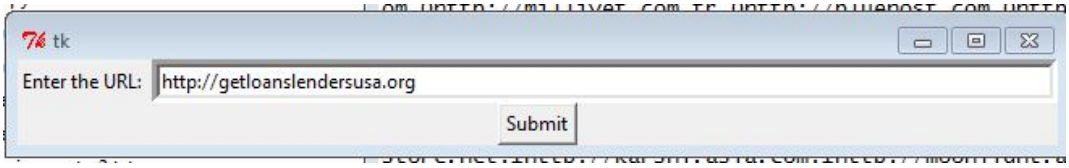
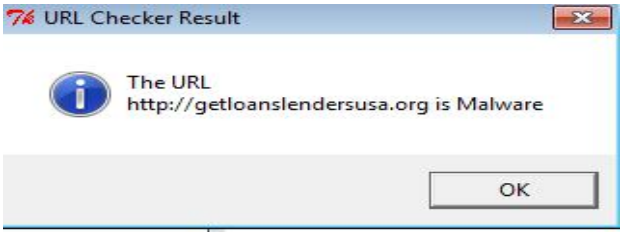


Fig3.4: training and testing using SVM algorithm

Now testing with different urls to find whether it is malicious or not As shown in Figure 3.5.



(a)



(b)

Fig3.5: Testing with URLs-(a) Entering url, (b) Testing url

IV. CONCLUSION

In the above paper we use the ICMC 2009 data set. We propose a method by utilizing the lexical features, Host based features (IP address, Packets, Token count), Web content based features. By using SVM algorithm we trained and classified datasets by optimizing SVM we get more accurate output then rest. Accuracy produced by Random forest is 95%, K-nearest Neighbor is 92%, and by using SVM we get accuracy of 97%.

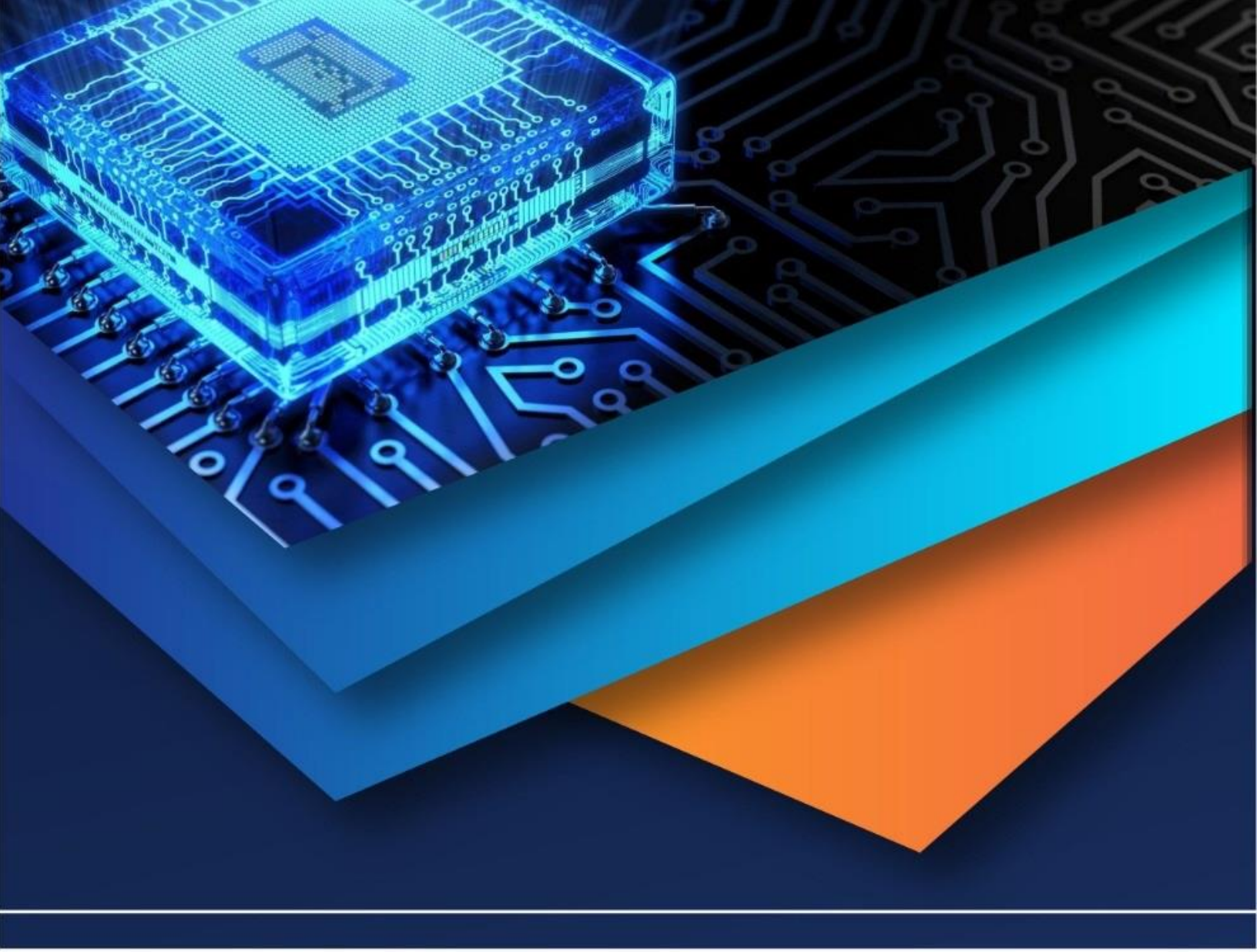
V. FUTURE SCOPE

In this project we use SVM algorithm with ICML 2009 data sets. In future we would like to carry our research to redirect of domain. The Diversion is not traced in this project. Hence malevolent website may contain diversion. In future we will use SVM redirection mechanism.

REFERENCES

[1] S. B. Rathod and T. M. Pattewar, "A comparative performance evaluation of content based spam and malicious URL detection in E-mail," 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, 2015,
[2] T. Zhang, H. Zhang and F. Gao, "A Malicious Advertising Detection Scheme Based on the Depth of URL Strategy," 2013 Sixth International Symposium on Computational Intelligence and Design, Hangzhou, 2013
[3] L. Fang, W. Bailing, H. Junheng, S. Yushan and W. Yuliang, "A proactive discovery and filtering solution on phishing websites," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015
[4] H. Sha, Q. Liu, Z. Zhou and C. Zheng, "GuidedTracker: Track the victims with access logs to finding malicious web pages," 2014 IEEE Global Communications Conference, Austin, TX, 2014,
[5] M. Akiyama, T. Yagi and M. Itoh, "Searching Structural Neighborhood of Malicious URLs to Improve Blacklisting," 2011 IEEE/IPSJ International Symposium on Applications and the Internet, Munich, Bavaria, 2011, pp. 1-10.
[6] J. Hong, “The state of phishing attacks,” Communications of the ACM, vol. 55, no. 1, pp. 74–81, 2012.
[7] P. Kolari, T. Finin, and A. Joshi, “Svms for the blogosphere:Blog identification and splog detection,” in Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, vol. 4, 2006.
[8] M. Dredze, K. Crammer, and F. Pereira, “Confidenceweighted linear classification,” in Proceedings of the 25th international conference on Machine learning. ACM, 2008,

[9] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong,and E. Nunge, "Protecting people from phishing: the design and evaluation of an embedded training email system," in Proceedings of the SIGCHI conference on Human factors in computing systems.
[10] J. S. Downs, M. Holbrook, and L. F. Cranor, "Behavioral response to phishing risk," in Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit.ACM, 2007
[11] M. Dredze, K. Crammer, and F. Pereira, "Confidenceweightedlinear classification," in Proceedings of the 25th international conference on Machine learning.
[12] Bergholz, J. Chang, G. Paaß, F. Reichartz, and S. Strobel, "Improved phishing detection using model-based features," in Proceedings of the Conference on Email and Anti-Spam (CEAS), 2008
[13] S. Garera, N. Provos, M. Chew, and A. Rubin, "A framework for detection and measurement of phishing attacks," in Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007.
[14] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," Proceedings of 17th Network and Distributed System Security Symposium.
[15] Moshchuk, T. Bragin, D. Deville, S. Gribble, and H. Levy,"Spyproxy: Execution-based detection of malicious web content,"in Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium
[16] R. Nagaonkar and U. L. Kulkarni, "Finding the malicious URLs using search engines," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 3692-3694.
[17] N. Provos, D. McNamee ET AL "The Ghost in the Browser: Analysis of Web based Malware," USENIX Workshop on Hot Topics in Understanding Botnet, 2007.
[18] N. Provos, P. Mavrommatis et al, "All Your iFrames Point to Us," USENIX,2008
[19] J. Nazario, "PhoneyC: A Virtual Client Honeypot," in USENIX Workshop on Large-Scale Exploits and Emergent Threats, 2009.
[20] M. A. Rajab, L. Ballard, "The Nocebo Effect on the Web: An Analysis of Fake Anti-Virus Distribution," in USENIX Workshop on LargeScale Exploits and Emergent Threats, 2010.
[21] M. Cova, C. Kruegel, and G. Vigna,"Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code," in International World Wide Web Conference (WWW), 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY