ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 6, Issue 4, pp: (292-298), Month: October - December 2018, Available at: www.researchpublish.com

Novel Design of Machine Learning for Malicious Software Analysis – Malicious URL Case Study

Dr.G.Anil Kumar[1], Dr.M.Upendra Kumar[2], Dr.Sesham Anand[3], Dr.D.Shravani[4]

[1] Professor and HOD of CSE Sridevi Women's Engg College, Affiliated to JNTUH Hyderabad TS India

anildeva@gmail.com

[2] Professor of CSE MVSR Engineering College, Affiliated to O.U. Hyderabad TS India

uppi_shravani@rediffmail..com

[3] Associate Professor of CSE MVSR Engineering College, Affiliated to O.U. Hyderabad TS India

anand_cse@mvsrec.edu.in

[4] Associate Professor of CSE Stanley College of Engineering and Technology for Women, Affiliated to O.U. Hyderabad TS India
sravani.mummadi@gmail.com

Abstract: **This research work proposes a novel and innovative idea of application of Machine Learning for malicious software analysis with a case study of malicious URL's implementations and validatings. Traditionally Data Mining and its associated tools were developed for Malware Detection. Also Data Mining and Machine Learning strategies were used in literature for Cyber security. Deep learning is also used for Malware Analysis. Machine learning techniques for Malware Detection includes: in supervised learning - Hidden Markov Model (HMM), Profile Hidden Markov Model (PHMM), Support Vector Machines (SVM) etc.; in unsupervised learning includes Principal Component Analysis (PCA), K-means etc. Machine learning for Web Mining includes strategies like: Web Structure Mining (Web Crawlers / Indexer/ Ranking – PageRank algorithm), Web Content Mining (Parsing), Natural Language Processing (Information Retrieval models –TF-IDF, Latent Semantic Analysis (LSA), Doc2Vec (word2wec), CBOW model), Post Processing (Latent Dirichlet allocation and Opinion Mining (sentiment analysis) etc.**

Keywords: **Malware Analysis and Detection, Machine Learning, Malicious URL classification.**

1. INTRODUCTION

By definition malware can be any category of malicious software which propagates into user systems bypassing authorization controls. Malware stands for malicious software. It is biggest and ever challenging threat for Internet today. As the number of websites increases exponentially day by day they become the common platform for the distribution of malware. The first step for a malware to evade a computer system is by downloading of files via Internet. Ones malware establish in presence in the system it continuously checks for the vulnerabilities of the operating systems and user applications then start performing unintended operations which eventually affects the overall performance of the system.

Malware can be extremely harmful because they have ability to infect exe files, data files, system files, partions of drives and can generate heavy traffic for initiating denial of service attacks. Whenever a user unknowingly a malware infected file then immediately that malware resides permanently in memory and initiates infection of all the files executed from then on wards. For operating systems vulnerabilities malware can even infect other system on networks.

Phishing is a harmful website with malicious intend containing malware to impersonate an authorized user and stealing sensitive information like banking details, credit card information and banking account credentials. In this era of social networking a phisher can get private information of user through technical deception. A phishing website does the trick of deserving user by providing this website with a look alike by all features as an authentic banking website and user cannot sense any difference and can give away a sensitive information to phishers. Phishing has taken online financial crimes to its peak and it is the most dangerous threat with in the web and can encourage web crime and cyber terrarium.

Machine learning uses computerized algorithms for processing and generation of complex models of data. These models data predictions may be accurate most of the time but some times rarely it may be inaccurate. Machine learning strategies explores data for choosing of efficient way to integrate information of representation of columns in data set, into a model which generalizes accuracy to data which was never seen before.

In literature we have thousands of machine learning algorithms for producing of data models. Few work in lightning fast, and others runs executing for even weeks for producing models. Some model are simple taking up few memory which others may take up MB's or even GB's of memory. Some run very fast to produce models for predictions on new data, others may require more computationally intensive executions. They may limitations that they work for only few types of data and may fail on other data / problems.

2. LITERATURE SURVEY

The exhaustive literature survey of important base papers is provided below for research problem definition and its solution.

[1] Provided detailed research methodology on Web Mining for Malware Detection. [2] Provided detailed implementations of machine learning for URL Classification and is the major source idea for our implementations and extensions. [3] Provided intelligent phishing URL detection using association rule mining. [4] Provides Malicious URL Detection and Identification. [5] Provides Malicious URL Filtering – A Big Data Application. [6] Provides Using Two Dimensional Hybrid Feature Dataset to Detect Malicious Executables. [7] Provides Data Mining Methods for Detection of New Malicious Executables. [8] Provides a Novel approach to protect against phishing attacks at client side using auto-updated white-list. [9] Provides Feature set identification for detecting suspicious URLs using Bayesian classification in social networks. [10] Provides A Study on Techniques for Proactively Identifying Malicious URLs. [11] Provides finding the malicious URLs using Search Engines. [12] Provides Fighting against phishing attacks: state of the art and future challenges. [13] Provides a practical approach for clustering large data flows of malicious URLs. [14] Provides Detecting Spam URLs in Social Media via Behavioral Analysis. [15] Provides Detecting Malicious URLs using Lexical Analysis. [16] Provides Lexical Feature Based Phishing URL Detection using Online Learning. [17] Provides Detecting Malicious Web Links and Identifying Their Attack Types. [18] Provides Identifying Suspicious URLs : An Application of Large-Scale Online Learning. [19] Provides Data Mining Tools for Malware Detection. [20]. Provides Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. [21] Improvement of classification features to increase phishing tweets detection accuracy. [22] Provides Defending against Phishing Attacks: Taxonomy of Methods, Current Issues and Future Directions. [23] Provides Malware Detection using DNS Records and Domain Name Features. [24] Provides Classification of unknown Web sites based on yearly changes of distribution information of malicious IP addresses. [25] Provides Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network: Case Study on Phishing Websites classification. [26] Provides Cyber Security Engineering. [27] Provides Introduction to Machine Learning with Applications in Information Security. [28] Provides Machine Learning and Security. [29] Provides Machine Learning for the Web. [30] Provides Data Mining and Machine Learning in Cybersecurity. [31] Provides Introduction to Artificial Intelligence for Security Professionals.

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 6, Issue 4, pp: (292-298), Month: October - December 2018, Available at: www.researchpublish.com

3. THEORETICAL ANALYSIS

Figure 1 provides the existing system in literature.

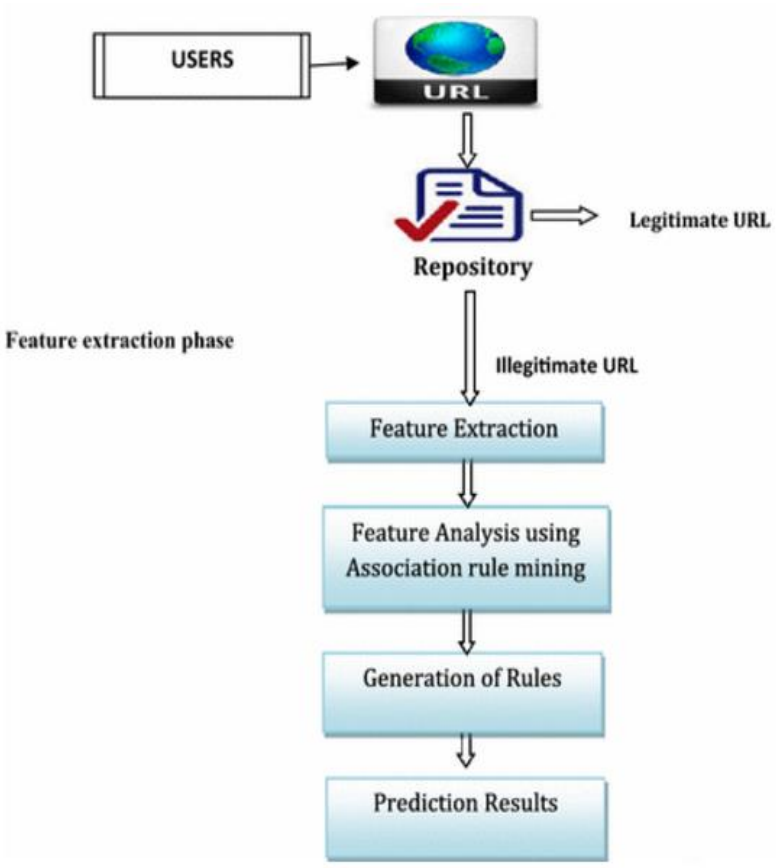

FIG 1: EXISTING SYSTEM

4. PROPOSED WORK

MALICIOUS URL Classification Case Study:

Malicious URLs (Uniform Resource Locator) are the major source of any distribution channels medium for broadcasting malware on Web. These malicious URLs will eventually become instrumental by providing partial or complete system control to the hackers. This will victimize user systems, which get easily infected by malware and, attackers can use these compromised systems for various Internet crimes exploiting the Dark side of Web, such as stealing user credentials, phishing, spamming, denial-of-service attacks etc. To detect such cybercrimes, the user systems should be very fast and have high precision with the new ability to analyze and detect new malware malicious content. This proposed Research gives introduction to various aspects which are associated with URL classification process, which will recognize whether the target website is a malicious or benign.

This research emphasis on efficient Data Mining classification algorithms to classify web url's.

This research focus on performance of features which are extracted from web url's. It employs authenticated current datasets. This new approach may detect malicious URL's significant, fast and better accuracy.

Malware is any kind of unwanted software that is installed without our consent on our computer. Viruses, worms, Trojan horses, bombs, spyware, adware are subgroups of malware.

Malware Detection is the strategy and methodology for detecting any latent and hidden Malware present in our systems. Data Mining classification techniques played a very important role in malware detection.

The application of this research is useful for all users using search engines and it will reduce financial losses at their individual levels based on severity of malware attack, for example recently seen Wannacry malware attack. Figure 2 provides the proposed system.

Feature Extraction strategies include: Lexical features, Link popularity features, Web content features, DNS(Domain Name System) features, DNS fluxiuess features, Network features, Domain Rank, Domain age, URL count, Longest Domain label, Dash count in Host name, Similar message count from a user, Similar message count from different user.

Datasets: Social Network datasets, Financial Sectors datasets, Phish Tank, DNS-BH, DMOZ Directory.

Classification Techniques: Decision Tree, Naïve Bayesian

Rule-Based, Support Vector Machine, K-Nearest Neighbor

Rough Set, Fuzzy set.

Performance Measurements: Accuracy, Relevance Analysis,

Speed, Robustness, Scalability, Interpretability

Exiting classifiers accuracy is less than 90% for different types of datasets. This new approach may detect malicious URL's significant, fast and better accuracy.

MALICIOUS URL Classification Case Study- implementations and validations:

Malicious web sites are the main central eccentric point of Internet criminal strategies. These web sites infect the visitor user systems with malware, viz. spam ads, phishing, etc. Traditional approaches to malicious URL problem are done manually by constructing lists which consists of list of malicious web pages URLs and its associated user systems which analyze the content and behavior of web site when it is visited. [2]

The major drawback of the above blacklisting strategy requires a very tedious work of searching the array table for its presence in the row. Moreover, this list is very huge keeping in view the number of websites on Internet also this list can't be update periodically as study shows web links grow on an hourly basis.

In the proposed methodology we are using design of machine learning strategies for classification of an URL for safe or unsafe dynamically with no need for downloading the webpage.

Algorithms used are Random Forest, Decision Trees, and Logistic Regression etc. This system now works only on simple text features of an URL viz. Presence of IP Address in Host Name / using the IP Address, Length of URL/Long URL to hide Suspicious Part, Domain length, , Presence of Security Sensitive words in URL, URL's having "@" Symbol, Redirecting using "//", Adding Prefix or Suffix Separated by (-) to the Domain, Sub Domain and Multi Sub Domain, Sub Domain and Multi Sub Domains, Domain Registration Length, The Existence of "HTTPS" Token in the Domain Part of the URL, Abnormal URL, Website Forwarding, Status Bar Customization, Using Pop-up Window, Age of Domain, PageRank, Google Index, Number of Links Pointing to Page etc.

This list can be expanded in the feature for addition of Server Host credentials like country code, date created, date updated etc., which will efficiently accuracy of the built classifier but delays the process of classifications by increasing latency time because it internally communicates with WHOIS server for these host features.

For implementations we used python language and for datasets we collected from phishtank.com.

With this proposed malware analysis and detection methodology we are exploiting key features which are robust for prediction of phishing or malicious websites using Lexical analysis of URL.

It is proved the vast literature that hackers always used very long URL in address bar for hiding malicious parts. Another trick of hacker to steal personal information is to use IP address in URL and to be more damaging IP address can be transformed into hexadecimal code. Also caution is required with @ in URL ignores by browser preceding @ and real address follows @ symbol. // path name redirects to other web links. As phishing site has very short life cycle they are never trustworthy in long run. Our classification machine learning algorithms distributes URL length to be either Malicious or benign. The features set taken as below [2]

1. Using the IP Address
2. URL's having "@" Symbol
3. Redirecting using "//"
4. Adding Prefix or Suffix Separated by (-) to the Domain
5. Sub Domain and Multi Sub Domains
6. Domain Registration Length
7. The Existence of "HTTPS" Token in the Domain Part of the URL
8. Abnormal URL
9. Website Forwarding
10. Status Bar Customization
11. Using Pop-up Window
12. PageRank
13. Google Index
14. Number of Links Pointing to Page

5. CONCLUSION

This research paper proposed a novel and innovative idea of application of Machine Learning for malicious software analysis with case study of malicious URL classification. Future work includes applying this innovative idea framework as technology transfer for Software Development industry for Secure Software Engineering for Systems and Software Assurance for Malware Analysis exploring research areas like: Code and Design Flaw Vulnerabilities, Malware Analysis Driven Use Cases. Distributed Systems involves high performance computing environments like cloud computing, Big Data etc... Design patterns and paradigm for scalable reliable services needs to explored using machine learning and implementations using docker, kubernetes, dcos etc... Further work involves implementations using HMM, PHMM, PCA, SVM, K-Means clustering for malware analysis and detection.

REFERENCES

[1] Shaik. Irfan Babu , Dr. M.V.P. Chandra Sekhar Rao , G. Nagi Reddy, "Research Methodology on Web Mining for Malware Detection" , IJCTT V12(4): 152-160, (2014),DOI:10.14445/22312803/IJCTT-V12P131

[2] https://github.com/surajr/URL-Classification/find/master (last accessed on 11.11.2018

[3] S. Carolin Jeeva, Elijah Blessing Rajsingh, "Intelligent phishing url detection using association rule mining", Springer Open, Human-centric Computing and Information Sciences, Jeeva and Rajsingh Hum.Cent.Comput.Inf.Sci.(2016), DOI: 10.1186/s 13673-016-0064-3

[4] Anjali B. Sayamber, Arati M. Dixit,"Malicious URL Detection and Identification", Internal Journal of Computer Applications(0975 – 8887) , volume 99 – No.17, (2014)

[5] Min-Sheng Lin, Chien-Yi Chiu, Yuh-Jye Lee and Hsing-Kuo Pao, "Malicious URL Filtering – A Big Data Application", IEEE International Conference on Big Data, Page No.589-596,(2013)

[6] Piyush AnastaRumao, "Using Two Dimensional Hybrid Feature Dataset to Detect Malicious Executables" , International Journal of Innovative Research in Computer and Communication Engineering, Vol.4, Issue 7,(2016), DOI: 10.15680/IJIRCCE.2016.0407158

[7] Matthew G. Schultz and Eleazar Eskin, Erez Zadok, Salvatore J. Stolfo, "Data Mining Methods for Detection of New Malicious Executables", Security and Privacy, S&P Proceedings. 2001 IEEE Symposium, SP'01, 38,(2001)

[8] Ankit Kumar Jain and B. B. Gupta, "A Novel approach to protect against phishing attacks at client side using auto-updated white-list", Springer Open, EURASIP Journal on Information Security, Jain and Gupta EURASIP Journal on Information Security,(2016),DOI. 10.1186/s 1335-0160034-3

[9] Chia-Mei Chen, D.J. Guan, Qun-Kai Su, "Feature set identification for detecting suspicious URLs uing Bayesian classification in social networks", ELSEVIER, Information Sciences 289 (2014),pg: 133-147.

[10] Asrian Stefan Popescu, Dragos Teodor Gavrilut, Dumitru Bogdan Prelipcean, "A Study on Techniques for Proactively Identifying Malicious URLs", International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, IEEE (2016), DOI: 10.1109/SYNASC.2015.40

[11] Amruta Rajeev Nagaonkar, Umesh L. Kulkarni, "Finding the malicious URLs using Search Engines", International Conference on Computing for Sustainable Global Development, IEEE(2016)

[12] B. B. Gupta, Aakanksha Tewari, Ankit Kumar Jain, Dharma P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges", Springer (2016).

[13] Adrian-Stefan Popescu, Dragos-Teodor Gavrilut, Daniel-Ionut Irimia, "A practical approach for clustering large data flows of malicious URLs", J Comput Virol Hack Tech , Springer (2016),pg:37-47, DOI. 10.1007/s11416-015-0239-x

[14] Cheng Cao and James Caverlee,"Detecting Spam URLs in Social Media via Behavioral Analysis", Springer International Publishing Switzerland (2015),pp. 703-714

[15] Mohammad Saiful Islam Mamun, Arash Habibi Lashkari, Natalia Stakhanova, Ali A. Ghorbani, "Detecting Malicious URLs using Lexical Analysis", Springer International Publishing(2016), pp. 467-482

[16] Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner,"Lexical Feature Based Phishing URL Detection using Online Learing", ACM (2010).

[17] Hyunsang Choi, Bin B. Zhu, Heejo Lee, "Detecting Malicious Web Links and Identifying Their Attack Types".

[18] ustin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker, "Identifying Suspicious URLs : An Application of Large-Scale Online Learning", International Conference on Machine Learning (2009).

[19] Mehedy Masud, Latifur Khan, and Bhavani Thuraisingham, "Data Mining Tools for Malware Detection", International Standard Book Number-13: 978-1-4665-1648-9, (2011).

[20] Justin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs", KDD (2009).

[21] SEOW WOOI LIEW,NOR FAZLIDA MOHD SANI, MOHD. TAUFIK ABDULLAH, RAZALI YAAKOB, MOHD YUNUS SHARUM, "IMPROVEMENT OF CLASSIFICATION FEATURES TO INCREASE PHISHING TWEETS DETECTION ACCURACY", Journal of Theoretical and Applied Information Technology(2018), E-ISSN: 1817-3195

[22] B. B. Gupta, Nalin A.G. Arachchilage, Konstantinos E. Psannis, "Defending against Phishing Attacks: Taxonomy of Methods, Current Issues and Future Directions"

[23] Khulood Al Messabi,Monther Aldwairi,Ayesha Al Yousif,Anoud Thoban,Fatna Belqasmi, "Malware Detection using DNS Records and Domain Name Features",ACM(2018), https://doi.org/10.1145/3231053.3231082

[24] Yoshitaka Nakamura,Shihori Kanazawa,Hiroshi Inamura,Osamu Takahashi, "Classification of unknown Web sites based on yearly changes of distribution information of malicious IP addresses",IEEE(2018)

[25] Grega Vrbancic,Iztok Fister Jr.,Vili Podgorelec,"Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network: Case Study on Phishing Websites Classification", ACM(2018),https://doi.org/10.1145/3227609.3227655

[26] Nancy R. Mead, Carol C. Woody, "Cyber Security Engineering", Pearson , ISBN : 978-93-325-8589-8

[27] Mark Stamp, "Introduction to Machine Learing with Applications in Information Security", CRC Press, ISBN: 978-1-1386-26782

[28] Clarence Chio, David Freeman, "Machine Learning and Security", O'REILLY ISBN:978-93-5213-693-3

[29] Andrea Isoni, "Machine Learning for the Web", PACKT ISBN:978-1-78588-660-7

[30] Sumeet Dua, Xian Du, "Data Mining and Machine Learning in Cybersecurity", CRC Press ISBN:978-1-4398-3942-3

[31] Cylance Data Science Team, "Intoduction to Artificial Intelligence for Securiy Professionals" CYLANCE Press ISBN: 978-0-9980169-0-0