Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

Design of Associate Content Based Classifier for Malicious URL Prediction by Rule Generation Algorithm

Vivekanadam Balasubramaniam,
Faculty of Computer Science and Multimedia,
Lincoln University College,
Kota Bharu, Malaysia.
E-mail: vivekanandam@lincoln.edu.my

Abstract- Recently, the internet is becoming as the most effective tool to interact with many foreign societies especially during COVID-19 pandemic. Moreover, the digital platform is increasing in many developing countries and at the same time, the chance of fraudulence is also increasing day by day. In the digital world, phishing assaults are emerging as the most common type of social engineering attack. Currently, many websites are targeting to acquire the confidential data, which is stored in websites. Recently, the classification techniques are employed to detect the phishing websites. Many tools are used for anti-phishing purposes; they are blacklist and antivirus software. The confidential data in a fake surrounding has intended the category of leaked data due to the action of attackers. In this scenario, machine learning method is observed as a very effective to classify the phishing and non-phishing web (Uniform Resource Locator) URLs. This classification struggles in classifying the leaked data content-based challenge. Therefore, the proposed algorithm is associated with the content-based classification method along with the rule-based generator algorithm. This research article integrates the content-based classification with a rule-based generator algorithm to improve the overall performance of the system. The updated public online repository called Mendeley dataset is used in the proposed research work. The proposed algorithm is used in 7k phishing and real websites content data for performing feature extraction. The extracted feature is then analyzed with our proposed algorithm to provide better prediction accuracy. Also, the proposed work has

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Information Technology & Digital World

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

concluded that, the associate algorithm has achieved better accuracy, when compared to other existing methods.

Keywords: *Deep learning, malicious website*

1. INTRODUCTION

Almost all people across the globe are connected through an internet source for many reasons. The information is stagnated in the open channel for a long time in the internet domain. All the business functionalities are surrounded by the internet banking sector from low level to high-level enterprises [1]. Figure 1 shows the strategy flow of the phisher.

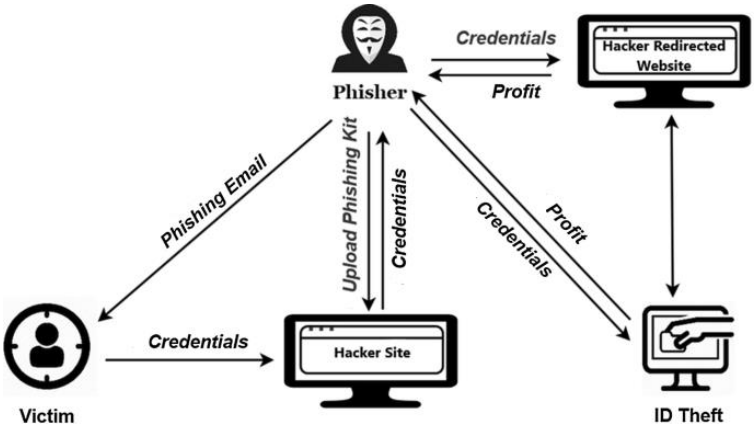

Figure 1 Strategy flow of Phisher

The online shopping website sometimes results in storing our credit card information and other similar confidential details. Many disadvantages exist in this, where users will not notice and be aware of it. Many phishing attackers are tracking the online shopping websites to get the structures of the websites [2]. On the other hand, the dangerous phishing websites is increasing day by day. They generally steal our confidential information, debit/credit card number, and passwords that are stored in our devices for future use [3]. The malicious URL performs much spam in our mail address, which is created by fraudsters to steal the confidential data from our devices. Figure 2 shows the attackers' controlled server.

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

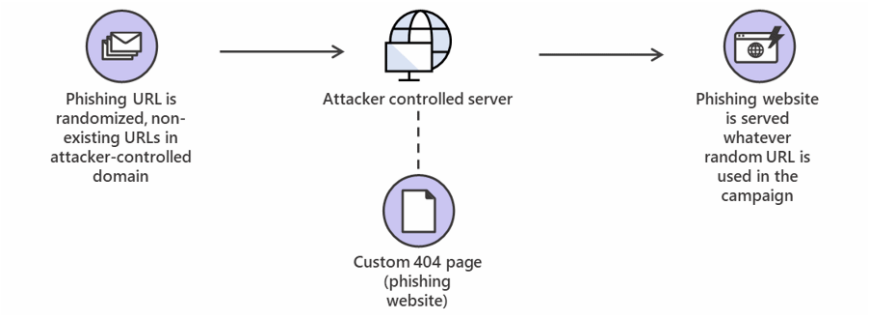

Figure 2 Attacker Controlled Server

The malicious URL consists of many attacking pins and malware to access the entire system without authentication. They can access our mobile webcam without our knowledge with the help of any malware or Trojan [4]. Many of the companies are taking several protective measures against these phishing attacks. But they couldn't succeed from attack due to the employee's continuous works or any unknown mistakes made in various websites, which is not customized [5]. The system security should be significant and reliable to prevent the entire server of the company from damage. The detection of faulty link and phishing attack prediction through any communication point is remaining as an essential and crucial requirement [6]. Many basic algorithms are practiced to detect phishing websites and protect the attacks before an attempt has been made by the attackers. But the phishing attackers will attack and damage the whole system in an intelligent manner. Figure 3 shows a simplified classification model for two different URLs [7].

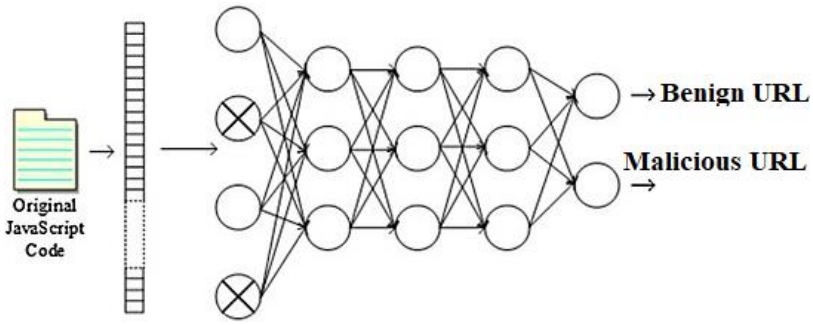

Figure 3 Simple Classifications by Java Script Code

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

Therefore, this research article is developing an algorithm to practice attack detection and study the behavior of attacking style in an intelligent way. This work can create another layer of security, which is used to detect the sites with false and fake features or information of the datasets. So the URL is a detecting point for the entry of the fraud website in the internet domain [8]. The detection of malicious URLs is an essential research topic in today's research and development. The solution meets the problems in the internet domain, which occur due to phishing attacks should be blacklisting. This process is based on the tedious task of analyzing the data with entry. This dataset is composed of a very large number of websites available on the internet [9]. This cannot be countable minute by minute to add with the dataset. These research papers are used to support in detecting the malicious harmful levels of the website and it contains virus through the URL. Therefore, the intelligence system is required to analyze the phishing website features and development structure [10]. The proposed algorithm is used to find the malicious websites from the internet domain. With the help of a machine learning algorithm, the model is trained to find the harmfulness of the website opened in our devices and also its browsing history.

2. ORGANIZATION OF THE RESEARCH

The research article is arranged as follows; section 3 deals with preliminaries of the proposed research work, section 4 gives the methodologies for classification function in order to detect the phishing URLs. Section 5 contains the discussion of results obtained by the proposed algorithm. Section 6 provides a conclusion and future task of the research article.

3. PRELIMINARIES

Recently, many algorithms are developed for predicting various malicious activities from phishing attacks. The estimation and feature collection from all the web pages are remaining as a very tedious process. Kan et al have proposed a machine-learning algorithm to predict the malicious website by using a classification method. A model has been trained for prediction

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

along with the help of lexical features obtained from page content present in the website during URL's host features [11]. Galera et al have proposed a machine learning algorithm with a logistic regression method for classification. They have selected 18 features for malicious URL identification and achieved 97.3% accuracy for their overall performance model [12]. McGrath et al conducted a test to analyze the phishing website with various methods. They were performed to categorize the non-phishing and phishing URLs [13]. Provos et al investigated the content-based feature extraction to detect malicious websites by using the machine learning method [14].

Moshchuk et al have investigated to detect the originality of the content from the downloaded material by using the antispyware tools [15]. Wang et al have believed non-machine learning approach to detect phishing content by utilizing the behavioral-based feature extraction method [16]. Choi et al presented machine learning methods to classify malicious URLs. They used kNN classifiers for achieving multiclass classification. They have extracted the features from the lexical characteristics. This includes the link volume structure, network traffic with DNS information. They have concluded that, the SVM classifier has highest performance accuracy to detect the malware function in websites. They introduced an obfuscation technique to detect malicious URLs with higher efficiency [17].

Ma et al discuss innovative techniques to detect malicious URLs by performing continuous assessment using online verification methods. They introduce logistic regression in a machine learning classifier along with a confidence weighted matrix. The features can be combined with web blacklist information for detection. Their proposed model is trained with the sliding window in the dataset. They solved dynamic malicious website problems with the help of a large and updated dataset to emphasize continuous training procedures. They have achieved very lower error rates in the detection process [18].

Fette et al have conducted a comparative study on machine learning algorithms and statistical methods for performing feature extraction from email structure to predict the irrelevant messages [19]. Bergholz et al improved the Fette et al papers by introducing text classification methods to predict the email content.. They achieved higher accuracy than the previously

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

proposed models [20]. Kolari et al introduced “bag of words” content ant it is determined for the malicious webpage of the URL [21].

4. METHODOLOGY

The detection of malicious websites from URL content is discussed briefly in this section along with various steps. Figure 4 shows the workflow of the proposed architecture.

Steps are in Phase I (Pre-Processed)

Step1: Data Pre-processing and the libraries like pandas, NumPy, yellow brick and sklearn are improted.

Step2: Import the dataset: The dataset was imported and named as “dataset”.

Step3: Construct a checking box to detect any checking missing and categorical values.

Step4: Split the dataset for training and testing with 80% and 20% respectively for obtaining more accurate results

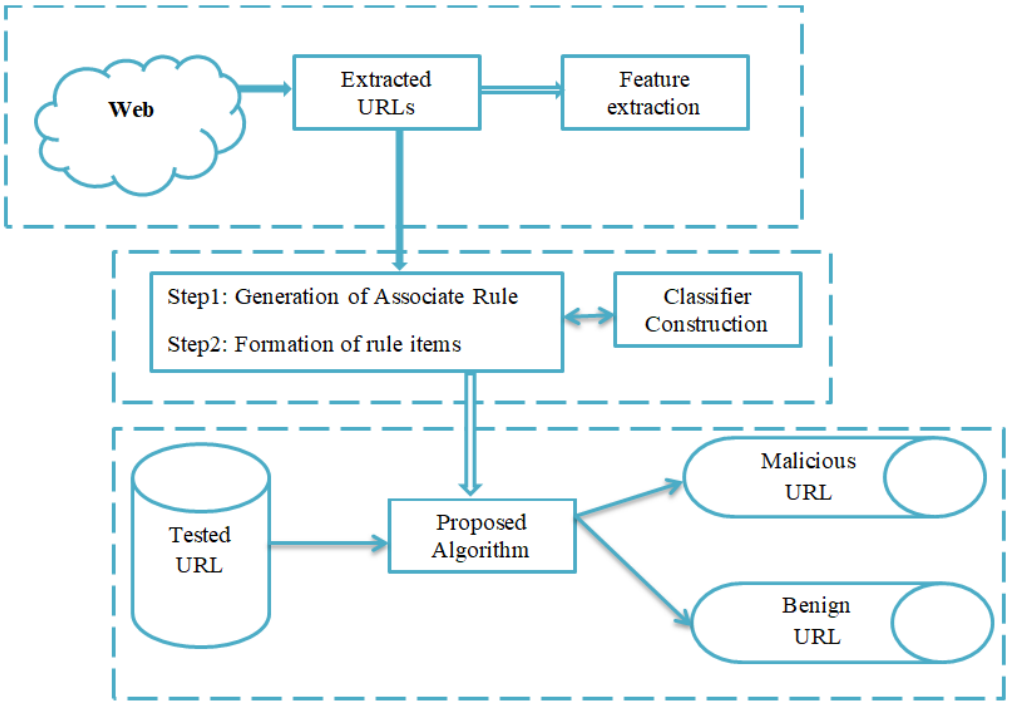

Figure 4 Proposed Architecture Workflow

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

Step5: Extract various URL features

The lexical and host-based features are extracted with textual properties present in the URLs. This can be used to distinguish the URL as malicious and benign. Special characters and sensitive words are also incorporated.

Step6: Webpage content and size based features extraction.

The line count and structure of the webpage can be measured to make it more accurate.

Step7: Java scripts code creation

This code includes doubtful patterns and functions to measure the malware distribution. The internal and external tags can be possible on the webpage. To evade detection, the obfuscation function has been incorporated because the attackers may attack via whitespace confuse automatic source code.

Steps are in Phase II (Classification)

Step8: Construct classification based on association techniques for malicious URLs detection.

Step9: Rule generator coding method is used to detect support measures items present in all transactions. This can create the confidence measures' threshold. The following formula can be derived,

$$Support(A) = \frac{Suppor\_count(A)}{Total\ Number\ of\ Transactions}$$

$$Confidence\ (A \rightarrow B) = \frac{Support\_count(A \cup B)}{Support\_count(A)}$$

Where, A and B feature; both are high then the URL can be classified as benign according to the model detection.

Step10: Construct the classifier builder. This rule-based classifier is sorting a cutback procedure to discard the redundant data from the dataset. Soon this makes optimization of the model after learned the dataset fully. So, this model is pruning the existing procedure for making an accurate classifier.

Step11: This algorithm traverses the database many times to obtain higher accuracy than the existing procedure. This optimizing procedure is used to access the data from the main memory

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

and it is performed with various codes of Java scripts that are relevant to detect the malicious URLs.

Step12: Selection of the rules for the classifier based on the sequences is sorted by using a rule-based algorithm.

Step 13: Optimized classification procedure

1. Integrated rules to find the instances in the database.
2. Instances are classified by using selected rules.
3. Ensure the instances for detecting nodes that are not matched with rule-based.

Step 14: If more relative errors arose, discard the above rule procedure and do instances changing procedure (which is called the automatic process) to improve the classification accuracy.

5. EXPERIMENTAL RESULTS AND EVALUATION

The benign URLs are authentic URL, which does not affect with the contagious website during the transaction. The malware URLs is categorized under malicious website, which is affected by attackers. This malware will steal the information from the device and it remains as dangerous software [22]. Both the benign and malicious 7K URLs are combined for our procedure. The proposed model uses accuracy, recall, and precision and confusion matrix to evaluate the performance. The formulas are defined as follows;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Confusion\ matrix = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

These confusion matrix metrics are used to evaluate the accuracy of the proposed model. The accuracy metrics are used for finding the ratio of correct prediction to the total samples in the proposed model. The proposed advanced algorithm is compared with many classifiers and tabulated in table 1. Also, this research work has observed the minimum support with 90% confidence, when it is having below 3% as shown in table 2.

Table 1 Performance Measure with Proposed Model

S.No	Methods	Accuracy	Precision	Recall	False Negative Rate
1	Random Forest	85%	81%	82.5%	8.47%
2	Decision Tree	86%	79%	80%	9.4%
3	SVM	87%	83%	81%	6.53%
4	Adaboost	86%	85%	84%	7.21%
5	Proposed Model	94.8%	90.1%	93%	2.40%

It can be concluded that, the classifier delivers a better performance. Also, if the minimum support is increased, the classifier performance gets decreased. The dataset has provided the features from labels of websites in order to acquire the details between malicious and benign URLs [23]. The user can be managed by using a phishing website and it includes errors in URLs. During this time, phishing URLs are used to steal the user information.

Table 2 Cross Validation by Confidence and Support Threshold

Support	Confidence (MinConf)				
	50%	60%	70%	90%	100%
1	90%	90%	85%	83%	79%
2	90%	90%	89.5%	80%	60%
3	88.9%	89.2	89.4	85%	64%

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

The training time is considered during the instances of objects found with a single event. Every deployed model is evaluated with the proposed framework. The proposed model achieves higher accuracy, recall, and precision. The time required for feature processing has been evaluated through a balanced dataset. Figure 5 shows the overall performance measure between the benchmark classifiers and the proposed ACBC algorithm.

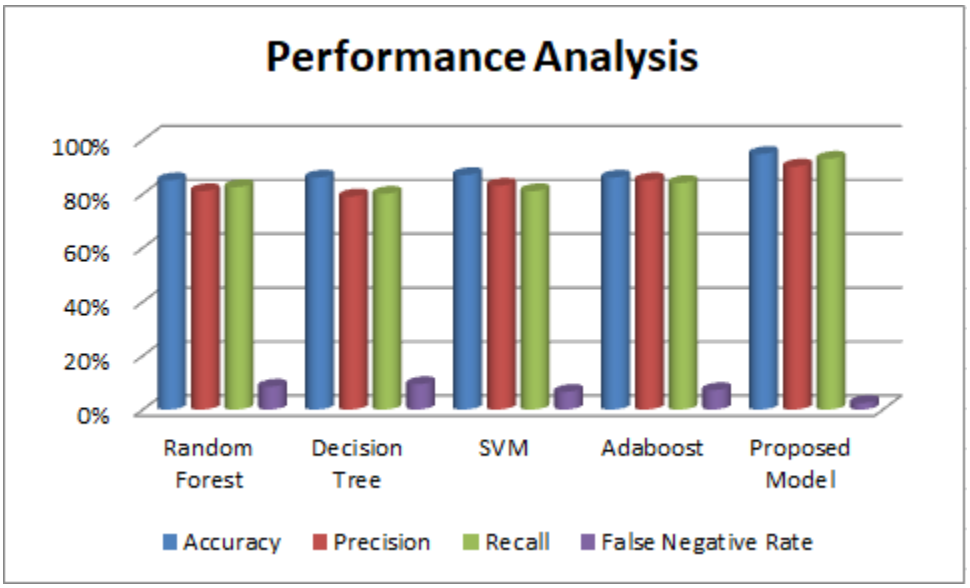

Figure 5 Overall Performance of Proposed System with Existing Classifiers

6. CONCLUSION

Thus, the proposed model is developed to detect malicious URLs with higher successive rates and accuracy. The proposed model is used to determine whether the given URL is malicious or benign with the help of lexical-based features. Therefore, the confidential information can be protected before it is stolen. The recent requirement consists of dynamic performance in the classifiers along with the associative function. The determination of malicious content in URLs allows a supervised learning domain. The proposed algorithm provides better accuracy and efficiency in dynamically detecting the malicious URLs. The experimental results are satisfied with the proposed algorithm in the architecture, when compared with various datasets, which includes the benchmark classification methods. Some of the limitations are present in the proposed study, which will be considered for future research

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

works; the dynamic modification in the dataset can be improved by the modified parameter to analyze the proposed model in the further amendment. Hopefully, this can provide improved defense in global sectors. Besides, the network configuration can be tuned by hyper parameters and its performance will be improved. The proposed algorithm will be extended to use in the browser, which we type and search URL; their intelligent decision can be taken by showing a message to proceed further. If the typed URL is correct, the browser page will run otherwise and closes automatically with the alert warning as further development in the proposed algorithm.

REFERENCES

[1] University of Waikato. WEKA. Available online: https://www.cs.waikato.ac.nz/ml/weka/ (accessed on 10 April 2020).

[2] Hahsler, M.; Johnson, I.; Kliegr, T.; Kuchaˇr, J. Associative Classification in R: Arc, arulesCBA, and rCBA. R J. 2019, 9, 254–267. [CrossRef]

[3] Jiˇrí, F.; Kliegr, T. Classification based on associations (CBA)—A performance analysis. In Proceedings of the CEURWorkshop Proceedings, Luxembourg, 20–26 September 2018; Volume 2204.

[4] Arntz, P. Explained: Domain Generating Algorithm. Available online: https://blog.malwarebytes.com/security-world/2016/12/explained-domain-generating-algorithm/ (accessed on 6 April 2020).

[5] Hadi, W.; Aburub, F.; Alhawari, S. A new fast associative classification algorithm for detecting phishing websites. Appl. Soft Comput. J. 2016. [CrossRef]

[6] Kim, S.; Kim, J.; Nam, S.; Kim, D. WebMon: ML- and YARA-based malicious webpage detection. Comput. Netw. 2018, 137, 119–131. [CrossRef]

[7] Li, Y.; Yang, Z.; Chen, X.; Yuan, H.; Liu,W. A stacking model using URL and HTML features for phishing webpage detection. Future Gener. Comput. Syst. 2019, 94, 27–39. [CrossRef]

[8] Google Safe Browsing. Available online: https://safebrowsing.google.com/ (accessed on 20 November 2019).

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

[9] Micro, T. 10 Scary Tricks Cybercriminals Use to Lure Unsuspecting Users. Available online: https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/10-scary-tricks cybercriminals-use-to-lure-unsuspecting-users (accessed on 20 January 2020).

[10] E. Uc¸ar, M. Incestas¸, and M. Ucar. A deep learning approach for detection of malicious urls. In Proc. of the 6th International Management Information Systems Conference (IMISC'19), Istanbul, Turkey, pages 12–20, October 2019.

[11] Kan, M.-Y. And Thi, H. O. N. . Fast webpage classification using url features. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM).(2005)

[12] Garera, S., Provos, N., Chew, M., And Rubin, A. D. . A Framework for Detection and measurement of phishing attacks. In Proceedings of the ACM Workshop on Rapid Malcode (WORM). Alexandria, VA.(2007).

[13] McGrath, D. K. And Gupta, M. . Behind phishing: An examination of phisher modi operandi. In Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET).(2008).

[14] Provos, N.,Mavrommatis, P.,Rajab,M. A., And Monrose, F. All your iFRAMEs point toUs. In Proceedings of the USENIX Security Symposium.(2008)

[15] Moshchuk, A., Bragin, T., Deville, D., Gribble, S. D., And Levy, H. M. SpyProxy: Execution-based detection of malicious web content. In Proceedings of the USENIX Security Symposium.(2007).

[16] Wang, Y.-M., Beck, D., Jiang, X.,Roussev, R.,Verbowski, C.,Chen, S., And King, S. Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities. In Proceedings of the Symposium on Network and Distributed System Security (NDSS).(2006)

[17] H. Choi, B. Zhu, and H. Lee. Detecting malicious web links and identifying their attack types. In Proc. of the 2nd USENIX Conference on Web Application Development (WebApps'11), Portland, Oregon, USA, page 11. USENIX, June 2011.

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021

Journal of Information Technology and Digital World (2021)
Vol. 03/ No. 01
Pages: 44-56
https://www.irojournals.com/itdw/
DOI: https://doi.org/10.36548/jitdw.2021.1.005

[18] J. Ma, L. Saul, S. Savage, and G. Voelker. Identifying suspicious urls: An application of large-scale online learning. In Proceedings of the 26th International Conference on Machine Learning (ICML'09), Montreal Quebec, Canada, pages 681–688. ACM, June 2009.

[19] Fette, I., Sadeh, N., And Tomasic, A. Learning to detect phishing emails. In Proceedings of the International World Wide Web Conference (WWW).(2007)

[20] Bergholz, A., Chang, J.-H., Paass, G., Reichartz, F., And Strobel, S. Improved Phishing Detection using Model-Based Features. In Proceedings of the Conference on Email and Anti-Spam (CEAS).(2008)

[21] Kolari, P., Finin, T., And Joshi, A. SVMs for the blogosphere: Blog identification and splog detection. In Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs.(2006)

[22] R. Verma and A. Das. What's in a url: Fast feature extraction and malicious url detection. In Proc. of the 3rd ACM on International Workshop on Security and Privacy Analytics (IWSPA'17), Scottsdale, Arizona, USA, pages 55–63. ACM, March 2017.

[23] J. Zhao, N. Wang, Q. Ma, and Z. Cheng. Classifying malicious urls using gated recurrent neural networks. In Proc. of the 12th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS'18), Kunibiki Messe, Matsue, Japan, volume 773 of Advances in Intelligent Systems and Computing, pages 385–394. Springer, Cham, June 2018.

ISSN: 2582-418X
Submitted: 14.03.2021
Revised: 7.04.2021
Accepted: 29.04.2021
Published: 18.05.2021