Covenant Journal of Informatics & Communication Technology. Vol. 5 No. 2, Dec, 2017

An Open Access Journal Available Online

Detecting Malicious and Compromised URLs in E-Mails Using Association Rule

Nureni Ayofe Azeez[1] & Emilia Anochirionye[2]

[1,2]Department of Computer Sciences,
Faculty of Science, University of Lagos, Nigeria
nazeez@unilag.edu.ng, emygreat1912@yahoo.com

Abstract: The rate of cybercrime is on the rise as more people embrace technology in their different spheres of live. Hackers are daily exploiting the anonymity and speed which the internet offers to lure unsuspecting victims into disclosing personal and confidential information through social engineering, phishing mails and sites and promises of great rewards which are never received. Thus resulting in great loss of property, finances, life, etc. and harm to their victims. This research work seeks to evaluate ways of protecting users from malicious Uniform Resource Locators (URLs) embedded in the emails they receive. The aim is to evaluate ways of identifying malicious URLs in emails by classifying them based on their lexical and hostname features. This study is conducted by extracting features from URLs sourced from phishing tank and DMOZ and adopting Association Rule of classification in building a URL classifier that analyzed extracted features of a URL and use it in predicting if it is malicious or not. 0.546 level of accuracy and an error rate of 0.484 was achieved as multiple URL features were employed in the classification process.

Keywords/Index Terms: Malicious, Association rule, URLs, Cybercrime, Hackers

1. Introduction

Information and Communication technology evolution has changed the way in which businesses are conducted all over the world. Prior to this era, messages where exchanged through courier and postal services, businesses where confided within the walls of an

organization and managing communication within and outside a business was pretty tedious. Contrarily, the advent of internet has increased the speed and automations with which businesses transact and communicate (Ayofe et. al., 2010). Irrespective of the institution, online visibility has become a key criteria for business survival and competitiveness within today's global business environment. It is however noted that online transactions are seriously being hampered across the globe because of insecurity (Azeez et. al., 2015). This is because, cybercriminals all over the globe are advancing on daily basis on their strategies to dupe and cause great financial loss on the internet users. This they achieve by sending fake and compromised URLs to internet users that perform online transactions (Azeez and Venter, 2013). Once such a user falls prey of their actions, they lose nearly if not all the money that is contained in their bank account. In an attempt to solve this challenge, researchers have adopted different data and machine learning algorithms to identify compromised URLs being used by phishers. Doing this has undoubtedly assisted internet users to identify which of the URLs sent to them is fake and phony. In this work, the authors proposed association rule for detecting malicious and compromised URLs in electronic mail. It our strong believe that this approach will supplement efforts made so far by researchers in this regard and specifically in the area of cybersecurity (Azeez and Ademolu, 2016).

1.2 Statement of problem

The benefits accruing from the internet evolution poses security challenges bordering around the preservation of intellectual, financial and personal information from fraudsters, who pose as legitimate internet users to steal valuable information from unsuspecting victims through vicious means such as malware, phishing mails, pharming, spams etc. Cybercrime has become a fast growing area of crime. Statistics from Forbe (Morgan, 2015) shows that cyber-attacks cost businesses as much as $400 to $500 billion a year excluding the cost of unreported cases. It is speculated that by 2017, the global cyber security market would skyrocket to $120.1 billion.(PWC, 2013)

To mitigate against the problem of users exposure to malicious links embedded in website or emails, the need to educate the populace on the risk of opening emails or attachments from unknown sender, clicking on links embedded in emails and ways of identifying a malicious link cannot be emphasized. However, this approach may not be sufficient as hackers rapidly adopts different ways of masking malicious URLs in legitimate emails and websites.

The alternative of matching suspected URLs against a blacklist which has been overly utilized, leaves the threat of a malicious URL going unidentified long before it is blacklisted. Currently, many researches have carried out researches and are still doing more in identifying malicious URLs using learning algorithms. This necessitates the need to evaluate the efficiency of this approach over the users' awareness campaigns and education on social engineering and malicious URL identification and adopting blacklist checks.

This research work centers on the "Identification of Malicious URL in Electronic Mails" using association rule and it aims to achieve the following:

- Implementation of a tool that extracts URLs from received electronic mails.
- Implementation of a URL analyzer that carries out analysis on URL based on some pre-defined features.
- Analyze the performance level in using each url features by using association rule

2. Literature Review

Previous works done in classifying URLs using machine learning algorithms adopted the methodology of classifying URLs based on the content of the website (Deri, 2015), which requires accessing and evaluating the contents of a website in order to ascertain its legitimacy.

Wardman and Warner used a technique that computes the similarity between the content files from potential and known phishing websites (Warner & Wardman, 2008). Ludl et al. classified phishing websites using features extracted from the main phishing webpage (Ludl, McAllister, Kirda, & Kruegel, 2007). This approach though more effective than traditional blacklist, poses a workload overhead on the classifier.

Work done by Soon and Jeffrey, evaluated the identification of a URL genuineness using Favicons (Soon & Jeffrey, 2014). This approach utilized google search-by-image API and semantic analysis in achieving 97.2% true positive in its classification.

Kan et. al. implemented URL classification by extracting features (Kan, 2005) in the URL and using them in the classifier. Initially features where extracted based on punctuation marks using tokens as the classifier's feature-set. Afterwards, researchers either used statistical (involved using statistical methods such as mean, variance calculation) to analyze the extracted information content (Kan, 2004) or brute-force approach to further segment URLs beyond punctuation marks. All possible sub-strings, n-grams in a URL are used as the classifier's feature-set in the brute-force approach (Iglesia, 2015).

Some of the URL features set utilized by researchers irrespective of their approach to URL classification using machine learning includes: lexical features (i.e URL features that URL string properties not considering host or page content) and External features (that is features that require querying a remote server) or a hybrid of the two features set to improve the performance of the classifier (Zhu, Lee, & Choi, 2001).

3. Proposed Framework

Process design is a technique used to describe the processes that transform data into useful information. It encompasses the flow of data through a system's process and /or logic, policies. This involves all the procedures to be implemented by a system's process. The flow chart below is used to explain the process flow of data as well as the policies that governed data flow and usage within the system.

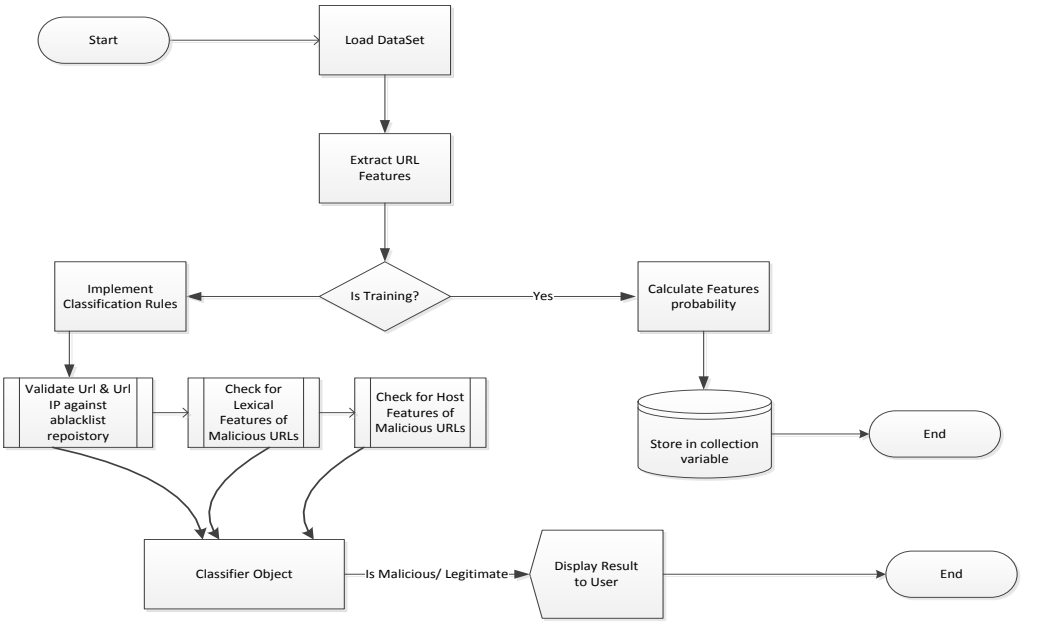

Figure 1: Process Flow Diagram

Algorithm 1: Association Algorithm

Algorithm: Association Rule

Rules:

Rule 1: if(URL IP is contained in IP blacklist)== true }class is malicious

Rule 2: If (URL is contained in blacklist database)==true } class is malicious

Rule 3: if (URL resolves to an IP Address) ==false } class is malicious

Rule 4: if (p(Malicious | Xo..Xn) > p(Legitimate | Xo..Xn)} class is malicious

Learning output Screenshot

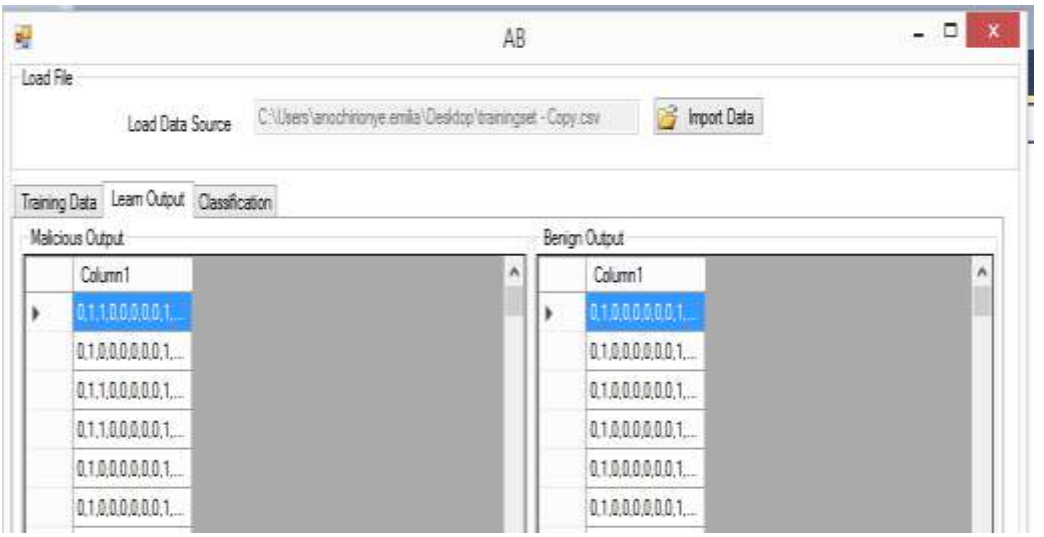

Figure 2: Training Data Result Screen

Classification Screenshot

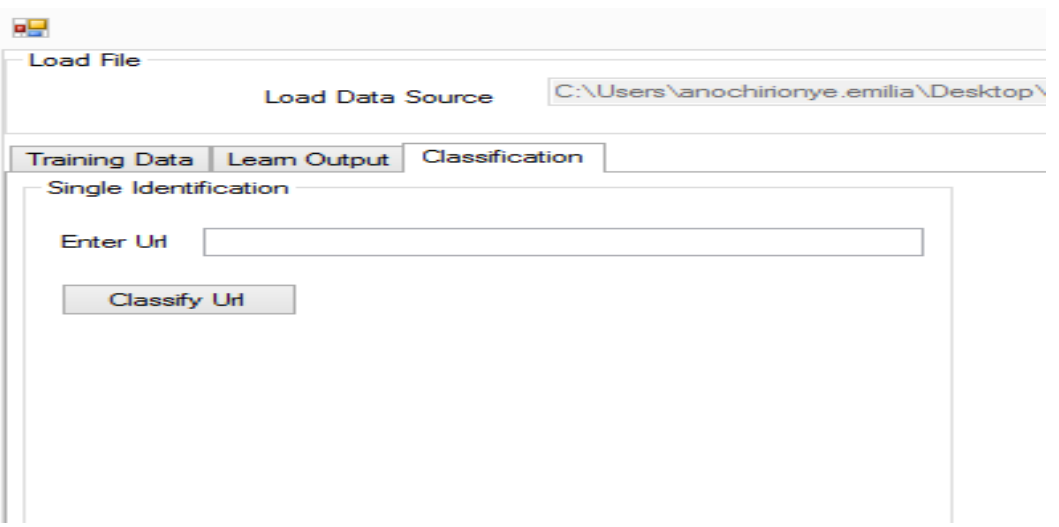

Figure 3: Single URL classification Screen

The screenshot for classifying URLs extracted from Emails is as shown below:

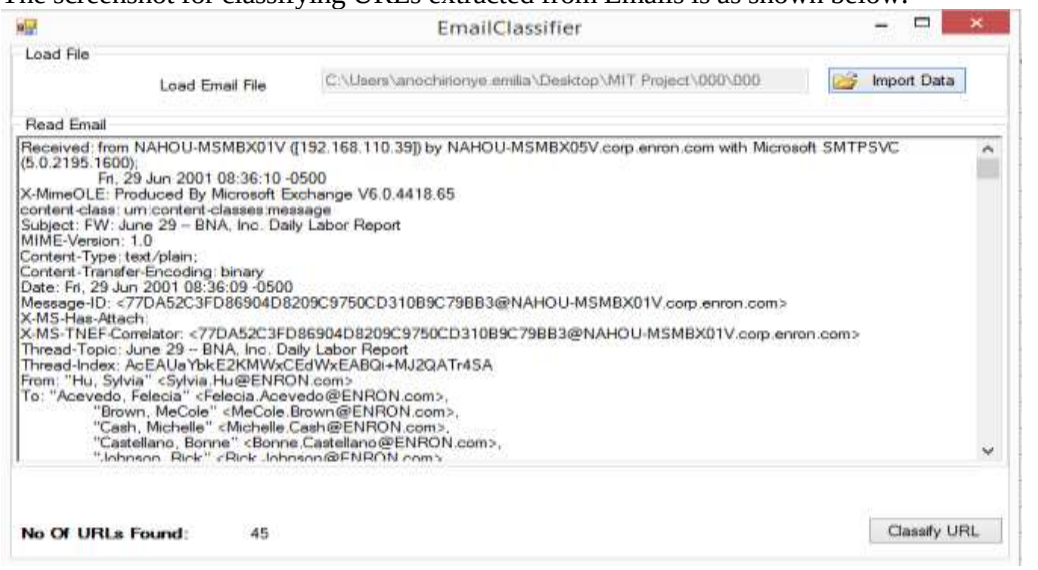

Figure 4: Classification Screen for urls in Emails

4. Output Design

This work has been designed to show users through screen outputs which can be exported out in excel format. The output screen are populated and presented to user mainly after features extraction or URL classification has been completed. The output report format adopted are "detailed reporting"; which one or more lines of output for each record processed is displayed. Each line of output printed is called a "Detail Line".

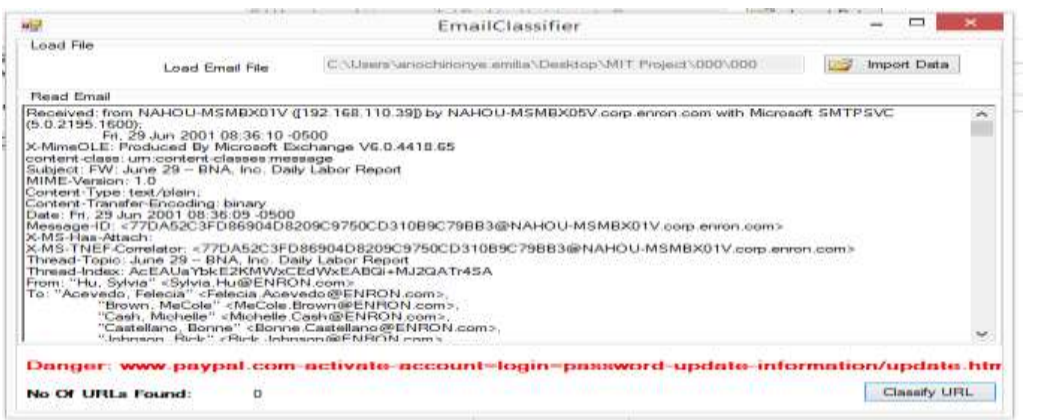

Figure 5: Email Test Data Result Screen

Figure 5 shows the screenshot obtained when some email dataset were used for the evaluation.

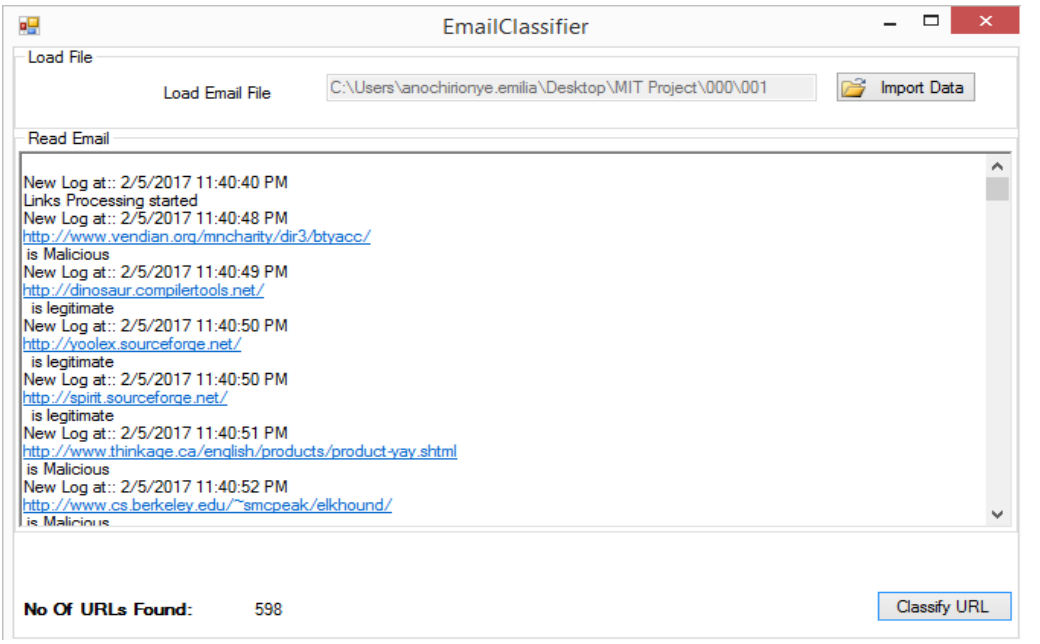

Figure 6: Email Test Data Result Screen

TABLE I: ANALYSIS OF FEATURES IN DATASETS

S/NO	Features	Frequency in Malicious dataset	Frequency in Legitimate dataset
1	Presence of URL Protocol (http/https) in the URL Path	4.0816%	0.0204%
2	Count of URL Resolved IP Address Not found in IP Blacklist	2.4490%	65.3061%
3	No of slashes in URL	42.8571	0.0204%
4	No of @ sign in domain	0.0204%	0.0204%
5	Count of IP in URL domain	0.0204%	0.0204%
6	Length of domain	0.0204%	0.0204%
7	Dots in domain	0.0408%	0.0204%
8	TLD found in TLDRepository	100%	97.9591%
9	No of Special Characters and keywords in URL	10.204%	0.0204%

Table I shows the corresponding values obtained for each of the features for the frequency of malicious and frequency of legitimate in the dataset. It should be recalled that there are nine (9) different features used for the evaluation. They are all listed in Table I.

5. Features Occurrence Evaluation

The probability of occurrence for each of the above mentioned features (See Table I) in the data classes for Malicious and Legitimate URLs was calculated in the training stage as shown in the Table I and results gotten from the analysis of these features in both Malicious and Legitimate datasets were used to build the predictive model used in detecting malicious URLs in the testing data based on the knowledge gained in the training stage on the rate of occurrences of the features in both the legitimate and malicious class.

i. Postulation of Decision Rule

From the review of the results obtained in the training stage, it was observed that most malicious URLs exhibited the presence of protocols (http/https) in their path, >5 slashes in the URL, unresolved IPs or IPs as their domain and the absence of TLD. Using association rule, which is a rule-based machine learning methodology that uses if/then statements that help uncover relationships between seemingly unrelated data in an information repository (Rouse, 2011), the Algorithm I was adopted in classifying the URLs in stage 3.

ii. Detecting Malicious Links within an Email

This is the last stage in the implementation of the URL classifier. It involves the extraction of over 40 URLs from a spam email dataset, extraction of the features predominant in malicious URLs and predicting the class (Malicious/ Legitimate) of each URL based on the knowledge acquired in the training stage. The association rules stated was used to evaluate the accuracy of the classifier.

6. Experimental Results

Results from the training stage on the probability of occurrence of each malicious URL features in the training dataset for the 2 classes (malicious and legitimate) is as shown in Figures 7,8,9 and 10. From the research, it is evident that validating URLs based on pre-defined features stated in the feature extraction stage will provide the following level of certainty for a dataset containing 50 URLs.

IP blacklist =24%

Number of slashes= 42.85%

Availability of special characters: 10.24%

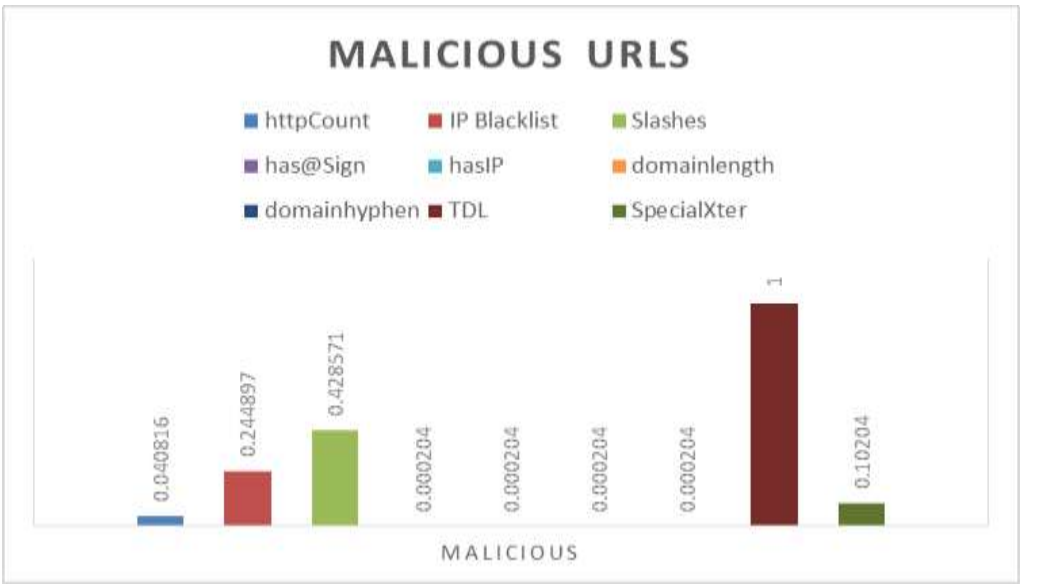

Figure 7: Malicious urls Feature Set Training Result for 50urls

Figure 7 shows the graphical representation of values obtained for malicious URLs with their corresponding features while Figure 8 depicts the graphical representation for legitimate URLs with their respective features.

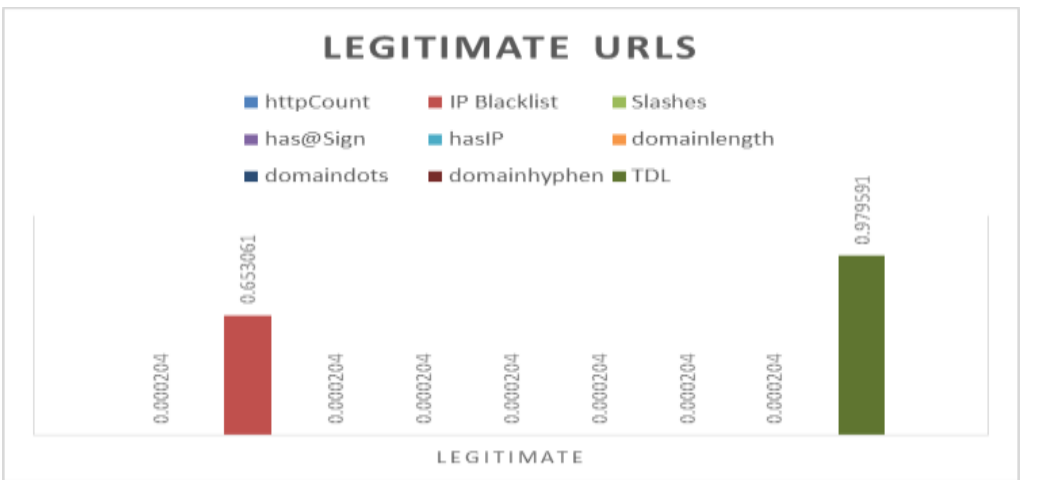

Figure 8: Legitimate Dataset Training Result for 50urls

While the presence of the Top Level Domain (TLD) in the repository, the valid TLDs gave the highest level of certainty.

Increasing the URL in the training dataset to 100, improved the certainty level for identifying malicious URLs using "domain length" and "domain dots" features from 0.000204 to 0.19 in legitimate URLs and 0.14 in malicious URLs.

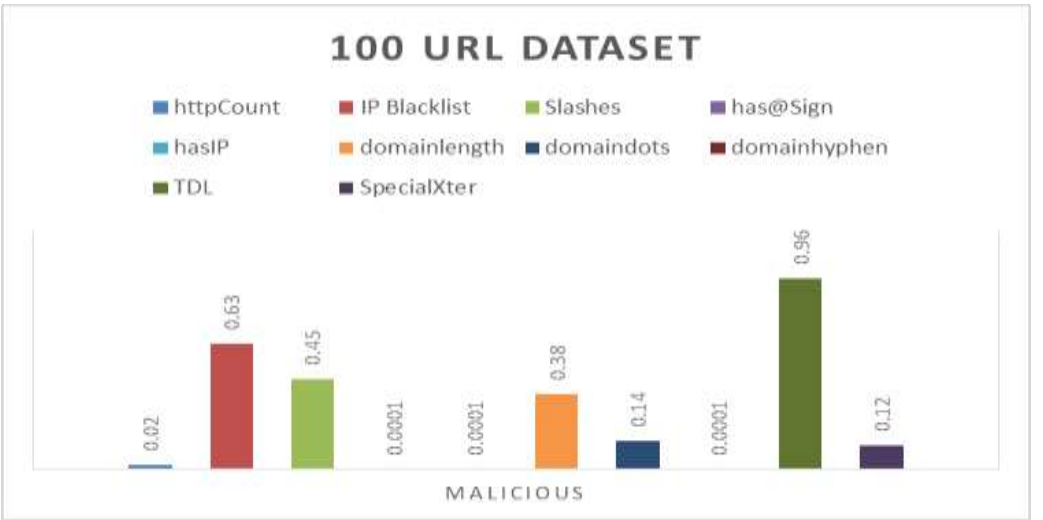

Figure 9: Malicious Dataset Training Result for 100urls

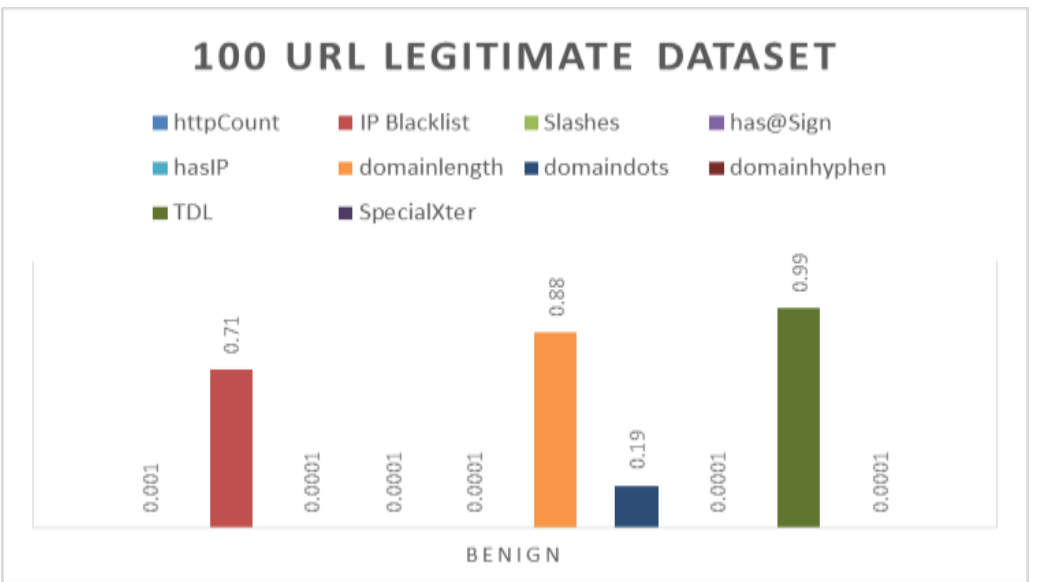

Figure 10: Legitimate urls Training Result for 100urls

URLs that failed detection based on Rules 1- 3 in the Association Rule Algorithm were moved to the classifier object and classification was done using the Bayes' rule by computing the following:

Posterior probability of a new URL(N) being legitimate = Prior Probability for Valid URLs * likelihood of N given a valid URL &

Posterior probability of a new URL(N) being malicious = Prior Probability for Malicious URLs * likelihood of N given a malicious URL.

Out of 100 malicious URLs and 100 legitimate URLs used in testing the classifier, the accuracy and error rates of the classifier were determined with the expression suggested by Damien (François, 2009):

Classification Accuracy = (TP + TN) / (TP + TN + FP + FN)

Error Rate = (FP + FN) / (TP + TN + FP + FN)

Where TP= True Positive, TN= True Negative, FP= False Positive and FN= False Negative.

The classifier developed in this research work, had 0.546 accuracy and an error rate of 0.484.

7. Conclusion

The rules implemented in the development of the URL classification in this research work can serve as a prototype for the development of a more robust solution which will help in identifying malicious URLS in email as well as most existing anti-virus solutions. Also, the continued advancement in the study and implementation of machine learning algorithms in the newly developed systems creates room for improved performance of the designed URL classifier in further works.

The major achievement of this project is the discovery of valid ways of:

Identifying malicious URLs in Electronic mails and

Analyzing URL features in classifying a URL as either legitimate or malicious.

Effort is ongoing to hybridize two different machine learning algorithms other than decision and association rule for the evaluation. Thereafter, a comparative assessment of the result with the association rule will be carried out.

Acknowledgement

The authors wish to acknowledge the efforts of anonymous referees for their valuable comments and helpful suggestions in shaping this paper into a publishable condition.

References

Ayofe, A.N, Adebayo, S.B, Ajetola, A.R, Abdulwahab, A.F (2010) "A framework for computer aided investigation of ATM fraud in Nigeria" International Journal of Soft Computing, Vol. 5, Issue 3 pp. 78-82

Azeez, N. A., & Ademolu, O. (2016). CyberProtector: Identifying Compromised URLs in Electronic Mails with Bayesian Classification. 2016 International Conference Computational Science and Computational Intelligence (CSCI) (pp. 959-965). Las Vegas, NV, USA: IEEE.

Azeez, N. A., & Iliyas, H. D. (2016). Implementation of a 4-tier cloud-based architecture for collaborative health care delivery. Nigerian Journal of Technological Development, 13(1), 17-25.

Azeez, N. A., & Venter, I. M. (2013). Towards ensuring scalability, interoperability and efficient

access control in a multi-domain grid-based environment. SAIEE Africa Research Journal, 104(2), 54-68.

Azeez, N. A., Iyamu, T., and Venter, I. M. (2011). Grid security loopholes with proposed countermeasures. In E. Gelenbe, R. Lent, and G. Sakellari (Ed.), 26th International Symposium on Computer and Information Sciences (pp. 411-418). London: Springer.

Azeez, N.A., and Lasisi, A. A. (2016). Empirical and Statistical Evaluation of the Effectiveness of Four Lossless Data Compression Algorithms. Nigerian Journal of Technological Development, Vol. 13, NO. 2, December 2016, 64-73.

Azeez, N.A, Olayinka, A.F, Fasina, E.P, Venter, I.M. (2015) "Evaluation of a flexible column-based access control security model for medical-based information" Journal of Computer Science and Its Application. Vol. 22, Issue 1, Pages 14-25

Azeez, N. A., and Babatope, A. B. (2016). AANtID: an alternative approach to network intrusion detection. The Journal of Computer Science and its Applications. An International Journal of the Nigeria Computer Society, 129-143.

Azeez, N.A and Otudor, A.E. (2016) "Modelling and Simulating Access Control in Wireless Ad-Hoc Networks" Fountain Journal of Natural and Applied Sciences. Vol 5(2), pp 18-30

Azeez, N.A Abidoye, A.P Adesina, A.O Agbele, K.K Venter, I.M Oyewole, A.S (2013) "Statistical Interpretations of the Turnaround Time Values for a scalable 3-tier grid-based Computing architecture" Computer Science & Telecommunications, Vol 39 (3), pp 67-75.

Zhu, H. L. (2001). Detecting Malicious Web Links and Identifying Their Attack Types. s.l., USENIX Conference on Web Application Development 2011 .

Brownlee, J., (2013). A Tour of Machine Learning Algorithms. [Online] Available at: http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ [Accessed 01 February 2017].

Ludl, S. M, (2007). On the effectiveness of techniques to detect phishing sites.. Switzerland, Conference on Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA).

Chandak, V., (2012). Parts of URL. [Online] Available at: https://www.virendrachandak.com/techtalk/parts-of-url/ [Accessed 01 February 2017].

François, D., (2009). Binary classification performances measure cheat sheet. Volume 10.

Grauschopf, S., (2016). A Simple Guide to Understanding URLs. [Online] Available at: https://www.thebalance.com/what-does-url-mean-897078 [Accessed 30 1 2017].

IBM, (2017). http://www.ibm.com. [Online]

Available at: http://www.ibm.com/support/knowledgecenter/SSGMCP_5.2.0/com.ibm.cics.ts.internet.doc/topics/dfhtl_uricomp.html [Accessed 30 January 2017].

IEEE, (2015). Large Scale Web-Content Classification. Programming and Systems (ISPS), 2015 12th International Symposium, Volume 10.1109/ISPS.2015.7244974, p. 15438528 .

Iglesia, T. A. (2015). URL-Based Web Page Classification:With n-Gram Language Models, s.l.: Springer International Publishing Switzerland 2015.

Kan, M., (2004). Web page classification without the web page. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters. s.l., ACM, pp. 262-263.

Kan, M. H. (2005). Fast webpage classification using url features.. The Proceedings of the 14th international conference on Information and knowledge, pp. 325-326.

Nureni , A. A., & Irwin, B. (2010). Cyber security: Challenges and the way forward. Computer Science & Telecommunications, 29, 56-69.

Morgan, S., (2015). The Business of Cybersecurity: 2015 Market Size, Cyber Crime, Employment, and Industry Statistics, s.l.: s.n.

PANDA, (2009). Malicious. [Online] Available at: http://www.pandasecurity.com/homeusers/security-info/213894/Malicious [Accessed 01 February 2017].

PWC, (2013). Cybercrime Event, s.l.: PWC Nigeria.

Rajsingh, S. C, (2016). Intelligent phishing url detection using association rule mining. s.l., Cross Mark.

Rouse, M., (2011). Definition: association rules (in data-mining). [Online] Available at: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining [Accessed 4 February 2017].

Rouse, M., (2016). machine learning. [Online] Available at: http://whatis.techtarget.com/definition/machine-learning [Accessed 01 February 2017].

Soon Fatt Choo, J. e. a., 2014. Phisdentity: Leverage website Favicom to offset Polymorphic Phishing Website. s.l., Conference Publishig Services.

Warner, B. W, (2008). Automating phishing website identification through deep MD5 matching. eCrime Researchers Summit, 29 October, pp. 1-8.

WHUK, (2007). types-of-url-absolute-and-relative. [Online] Available at: https://www.webhosting.uk.com/blog/types-of-url-absolute-and-relative/ [Accessed 31 January 2017].