ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Malicious URL Detection Using Machine Learning Techniques

R.Jeena[1], G.preethi[2], A.Praveena[3], A.preethi[4]

Assistant Professor, Department of IT, Panimalar Institute of Technology, Chennai, India[1]

UG Scholar, Department of IT, Panimalar Institute of Technology, Chennai, India[2,3&4]

ABSTRACT: In this paper, we discuss the importance of integrating Machine learning techniques in the process of detecting malicious URL. With the rise of smart phone usage, many transactions are conducted online. Even sensitive transactions are going online. A static approach to security is always a threat. By introducing a dynamic approach, we can greatly reduce the phishing attacks. We have used supervised learning algorithms like SVM which gives an accuracy of 0.81 and F1 score of 0.74. These algorithms dynamically adds malicious URL once detected. This prevents attacks caused by carelessness of end users.

KEYWORDS: SVM(support vector machine),URL(uniform Resource Locator), Blacklisting, F1 score

I. INTRODUCTION

Every business regardless of its sector has its website. Nearly 30,000 websites are hacked everyday[1]. Though there are many security products to prevent the suspicious attempt by attackers, hackers are coming up with new ways to breach into the system. Not all the users are sufficiently aware when it comes to Security. Still people are the weakest link in security. Since Phishing attack is a social engineer attack[2], people end up giving their confidential information to the hackers. Especially financial Fraud tricks many unsuspected users to give their credit card details. There are many medium that are used to stick the malicious URL to. Email is the most prominent way used by the attackers. In 2017, it is reported that spam contributed 56% of the total email traffic[3], which amounts to half of the total traffic. Attackers embed the malicious URL in a document or an image that makes the users to open it eventually leading to a malicious site[7]. Information that users enter on this site gets stored on the remote database being controlled by the attackers. The phishing website exactly looks the same as an original website. So the best way to prevent ending up at such site is by integrating supervised learning algorithms such as SVM(Support Vector Machine) and CNN(Convolutional Neural Network). These algorithms helps to detect the authenticity of the website. It performs dynamically by adding malicious URL once captured. In the classic approach, list of malicious URL are stored in the database. On requesting a website, the URL is compared only with the already stored URL[3] but this approach fails to find newly designed phishing sites. With the implementation of Machine learning algorithms newly emerging Phishing sites could be detected and reported to the user. It also adds the malicious site to the database immediately.

II. RELATED WORK

The previous work on malicious URL detection is broadly categorized into two analysis techniques namely Active analysis and Passive Analysis[6]. Active analysis includes DNS detection[3], website content analysis and other such applications. Passive analysis is further consists of three approaches. They are rule based matching, machine learning approach and graph based approach. Rule based approach detects fraudulent domains[1] by matching the requested URL with the list of domains in the already prepared list or it spots out malicious domain by applying specific rules. The machine learning approach is the prominent approach used widely. It is the reliable approach compared to other approaches. It has a high predictability rate with high accuracy. It utilizes host based features and it lexical features to find whether the entered URL is authentic or not[1]. The pattern of a normal authentic URL is compared with the pattern

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

of the entered one. It gives attention to finer details of the domain. The graph based approach is recently proposed approach that outperforms the other two approaches. It collects information about the website from various sources like WHOIS[2] and applies efficient algorithms, in this case which is Belief Propagation(BP)[2]. Additionally path based algorithms are also used to derive the relationship between the domains. While comparing Artificial Neural Network (ANN) approach with static approaches[3] it is concluded that the ANN approach outperforms all other approaches and it is much more efficient.
It gave the best results with accuracy as high as 95.09%. This approach greatly reduced wrong positive results.

III. PROPOSED SYSTEM

We propose Machine Learning algorithms specifically supervised learning algorithms to find the malicious URL. The supervised learning algorithms include Support Vector Machine (SVM) and Convolutional Neural Network (CNN). As in the case of supervised learning algorithm, we provide the system with a set of inputs. These inputs consists of the already detected URL and a set of features that usually makes the malicious URL. We include nearly 1000 malicious websites collected from reputed sources like Alexa ranking. We also go through the malware database to pick up the recently found malicious sites. To make this approach even more robust, we used logistic regression which is almost similar to linear regression. Here logistic regression is used for classification and also to predict value. It uses linear equation for prediction. It sets the range of the linear equation output to [0,1].

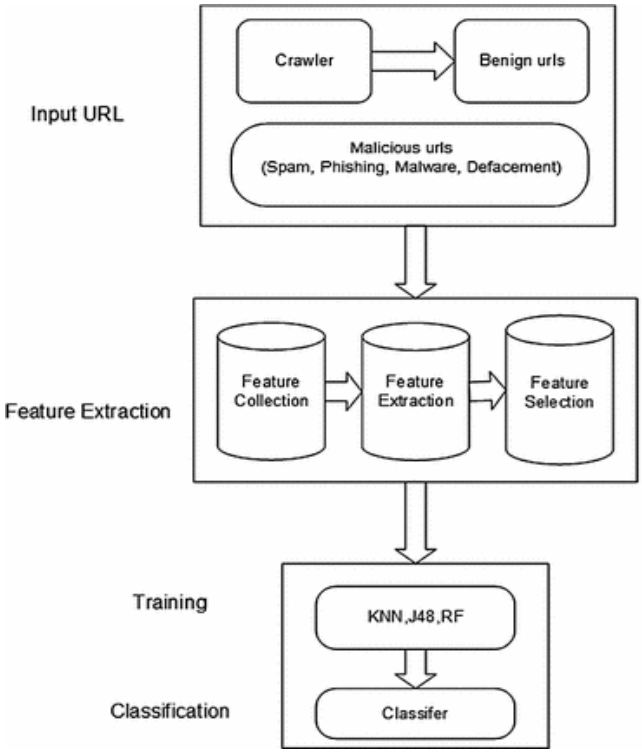

Fig. 1 Block Diagram of Proposed System

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

TABLE I: Description of data set

Data set	Training	Testing
Data set 1	60,000	30,101
Data set 2	-	26,000

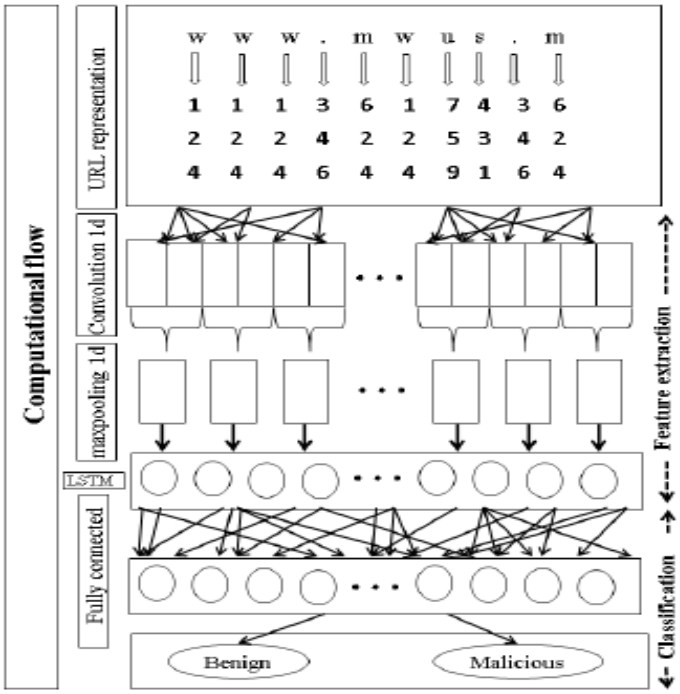

Fig. 2 An intuitive overview of proposed CNN-LSTM model.

IV.EXPERIMENTAL RESULTS

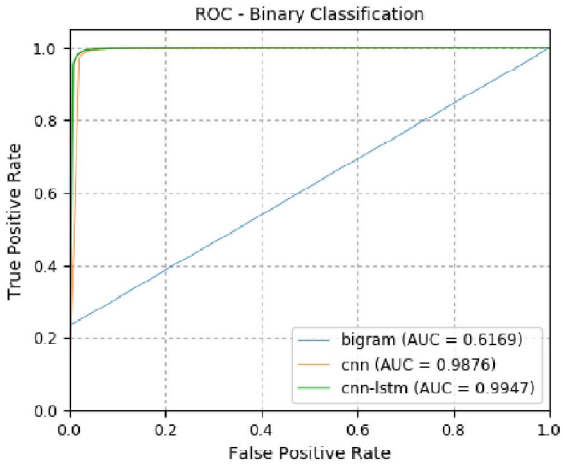

Fig. 3 ROC curve of bigram CNN and CNN-LSTM classifier

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

With logistic regression other models such as bigram, CNN and CNN LSTM[9] are used. These models are applied on two datasets namely Dataset 1 and Dataset 2. By comparing true positive rate and false positive rate, an ROC curve[4] is drawn which you can see in the diagram. The threshold range is [0.0-1.0]. The complete report on statistical measures of Dataset 1 and Dataset 2 is obtained. The result proves that deep learning model comparatively works well. All other models work less for Dataset 1 and Dataset 2. Thus deep learning layers[13] at character level is a powerful approach for malicious URL detection.

V.CONCLUSION

Cyber security is an ever growing field. With new vulnerabilities and security exploits popping up each day, security should be made robust. Awareness on cyber security shouldn't only be bound to security professionals rather it Should be extended also to the developers and the end users. Integrating Machine learning techniques will bring a drastic improvement to the overall security and may greatly reduce security breaches and Phishing rates. It is conspicuous from the experimental results that deep learning techniques outperforms other approaches. Supervised learning algorithms is a potentially significant step in discovering malicious URL. We hope to bring forth significant techniques to close the space of feasible attacks.

REFERENCES

[1]A. Sirageldin, B. B. Baharudin, and L. T. Jung, "Malicious web page detection: A machine learning approach," in *Advances in Computer Science and its Applications*. Springer, 2014, pp. 217–224.
[2] Y. Kim, "Convolutional neural networks for sentence classification,"*arrive preprint arXiv:1408.5882*, 2014.
[3] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.
[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[5] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget:Continual prediction with lstm," 1999.
[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[7] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.
[8] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.
[9] F. Chollet, "Keras (2015)," *URL http://keras. io*, 2017.
[10] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[11] K. P. S. Vinayakumar R and P. Poornachandran, "Evaluating deep learning approaches to characterize and classify malicious urls," *Journal of Intelligent Fuzzy Systems*, vol. 34, no. 3, pp. 1333–1343, 2018.
[12] R. Moazzezi, "Change-based population coding," Ph.D. dissertation, UCL (University College London), 2011.
[13] A. K. Jain and B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," *EURASIP Journal on Information Security*, vol. 2016, no. 1, p. 9, 2016.
[14] M.-Y. Kan and H. O. N. Thi, "Fast webpage classification using url features," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 325–326.
[15] B. Liang, J. Huang, F. Liu, D. Wang, D. Dong, and Z. Liang, "Malicious web pages detection based on abnormal visibility recognition," in *EBusiness and Information System Security, 2009. EBISS'09. International Conference on*. IEEE, 2009, pp. 1–5.
[16] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in *Proceedings of the 2007 ACM workshop on Recurring malcode*. ACM, 2007, pp. 1–8.