IJIRMET
ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering and Technology
Vol. 7, Issue 3, March 2022

An Effective Approach To Detect Phishing Websites Using Random Forest Algorithm

[1] Ms.R.Srividya,[2]B.Aishwarya,[3] D.Monasakthi,[4] S.Snegha

[1] Assistant Professor,[2][3][4] B.E-Final Year Department of Computer

[1][2][3][4] Department of Computer Science Engineering, Adhiparasakthi Engineering College , Melmaruvathur,Kanchipuram, Tamil Nadu.

Abstract: ***In recent years, advancements in Internet and cloud technologies have led to a significant increase in electronic trading in which consumers make online purchases and transactions. This growth leads to unauthorized access to users' sensitive information and damages the resources of an enterprise. Phishing is one of the familiar attacks that trick users to access malicious content and gain their information. In terms of website interface and uniform resource locator (URL), most phishing webpages look identical to the actual webpages. Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested. However, due to inefficient security technologies, there is an exponential increase in the number of victims. The anonymous and uncontrollable framework of the Internet is more vulnerable to phishing attacks. Existing research works show that the performance of the phishing detection system is limited. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the another proposed a URL detection technique based on machine learning approaches.A recurrent neural network method is employed to detect phising URL.Researcher evaluvated the proposed method with 7900 malicious and 5800 legitimate sites,respectively.The experiments outcome shows that the proposed method performance is better than the recent approaches in malicious URL detection.***

Keywords-Logistic Regression,Multinomial Naïve Bayes,XG Boost.

1. INTRODUCTION

Consumers have lost billions of dollars each year as a result of phishing operations. Refers to thieves' tricks for obtaining private information from a group of unwitting Internet users. Fraudsters obtain personal and financial account information such as usernames and passwords using fake email and phishing software to steal sensitive information. This research examines strategies for detecting phishing Web sites using machine learning techniques to analyze various aspects of benign and phishing URLs. It investigates how linguistic cues, host features, and page significance attributes are used to identify phishing sites. It examines several machine learning techniques for feature assessment in order to acquire a better understanding of the structure of URLs that facilitate phishing. The fine-tuned parameters aid in the selection of the most appropriate machine learning method for distinguishing between phishing and benign sites. Criminals that seek to steal sensitive information first establish illegal duplicates of legitimate websites and e-mail accounts, frequently from financial institutions or other companies that deal with financial data. The e-mail will be made up of real firm logos and slogans. One of the reasons for the rapid growth of the internet as a means of communication is that it allows the misuse of trademarks, brand names, and other corporate identities that consumers rely on as verification processes. "Spoof" e-mails are sent to many people in order make them involved in the criminal deception. Consumers are paid on a fraudulent website that appears to come from the real company when these emails are opened or when a link is clicked on the email

II. LITERATURE REVIEW

According to Erzhou Zhu (2018), phishers typically put up a false website where victims were tricked into providing passwords and perceptive information[1]. As a result, it's critical to detect rogue websites before they cause any harm to their victims. This study proposes a new method based on deep reinforcement to model and detects malicious URLs, fueled by the dynamic nature of criminal websites to steal sensitive information[2]. The suggested model may learn the properties related to phishing website identification by accommodating the dynamic behavior of phreaking websites[3]. The use of various types of machine learning algorithms for the problem of classification, particularly security and virus detection, has piqued the research community's interest in recent years[4]. Deep learning algorithms have also opened a new chapter on pattern recognition and artificial intelligence with the growth of processing capacity[5]

. As a result, these powerful learning algorithms may now be used to solve a wide range of categorization, decision, and automation challenges[6]. When a high number of characteristics are included in the computation, deep learning-based techniques are very effective. Because algorithms based on reinforcement learning may estimate solutions (i.e., action) based on stochastic transformations and the rewards of selecting that state action, the proposed method is robust and flexible[7]. By examining the given URLs, this project proposes a deep reinforcement learning-based model to find phishing websites[8]. The model adapts to modify the URL structure on its own. The traditional classification challenge is exemplified by the problem of recognizing phishing websites[9]. To handle this categorization challenge, a reinforcement learning model based on deep neural networks is constructed[10]. The problem of an "agent" performing an action that is entrenched on "trial and error" through interactions with an uncharted "environment" that offers a response in the form of numerical "rewards" is defined as the adaptive learning paradigm[11]. Other deep learning-based 4 algorithms, such as LSTM, should be investigated for the challenge presented in this research[12]. This classifier can also be used to solve other binary classification problems, such as detecting Webspam and the presence of hostile bots in the network[13]. Because the classifier in the RL-based technique is more flexible, it can be used to address a variety of privacy and security concerns in wearable devices[14].

To find criminal websites and its objective, SeenaThomas(2017) recommended extracting features from URLs and webpage links[17]. The matrix element is made up of basic links to the webpage of a given URL, in addition to the basic URL properties provided, such as length, suspicious characters, and a number of dots. In addition, statistical features such as mean, average, and variance are retrieved from each column of the feature matrix[21]. The given URL, links, and content on its web page, such as title and text content, are also used to extract dictionary features[24]. In order to detect the crime of identity theft, a number of machine learning models were tested, with the Deep Forest model showing competitive performance[25]. In particular, to discover hacker targets, an adequate technique based on search operators through search engines was devised. This method is rapid, however, it does not identify newly constructed phishing URLs[26]. Heuristic-based algorithms extract textual features, which can recognize newly created URLs, to detect phishing websites. However, some textual components extracted from webpage content cannot be used to identify phishing websites in other languages[27]. According to some scholars, analogy- based methods should be used to compare the similarities between the actually given web pages under attack, i.e. the target of identity theft, which should be identified in advance. The criminal intentions of stealing sensitive information can now be seen automatically[28]. However, this approach is slow because it requires finding and analyzing a large number of web pages in order to improve the parasite community[29]. Using URL representation in vector form, Deep Forest plus a variety of current machine learning models, such as GBDT and XG Boost, may be applied to detect sensitive identity theft. [30]. The proposed method works with websites written in a
variety of languages. The obtained features can be employed by a variety of classification methods, with DF outperforming the competition[31].

III. PROPOSED METHODOLOGY

The model is preprocessed in the proposed system, the words are tokenized, and stemming is performed. Data Processing is the process of converting or encoding data for easy machine transfer. In other words, the algorithm can now easily define data features. We must vectorize our URLs now that we have the data. Because some words in URLs are more essential than others, such as "virus," ".exe," and so on, the model employs Count Vectorizer and tokenizer to aggregate words. Let's make a vector representation of the URLs. A tokenizer that separates a string using a regular expression that matches either the tokens or the separators between tokens is known as a regular expression tokenizer. A regex pattern is a particular language for representing general text, numbers, or symbols in order to extract texts that match the pattern. 's+' is a simple example.. The method will sync at least one or more gaps if you add a '+' at the end.. In the world, stemming is crucial. Queries and Internet search engines both use stemming. The Fast Api framework is used for deployment. Fast API is a web framework for constructing APIs with Python 3.6+ and standard Python type hints that is current and fast (high-performance). The following are the main characteristics: Fast: Extremely fast, comparable to NodeJS and Go (thanks to Starlette and Pydantic). One of the quickest Python frameworks on the market. The UI is provided using FastAPI by loading the machine learning model into it. The architectural flow is shown in fig.1..

A. Advantages of proposed system

- User Interface is provided
- Model is trained using many features
- High level of accuracy

B. Logistic regression

A statistical strategy for predicting binary classes is logistic regression. The outcome or target variable is a binary variable. The term dichotomous refers to the fact that there are only two potential classes. It can, for example, be utilized to solve cancer detection issues. It calculates the likelihood of an event occurring.

ISSN (Online): 2456-0448
International Journal of Innovative Research in Management, Engineering, and Technology
Vol. 7, Issue 3, March 2022

C. XG Boost

Extreme Gradient Boosting is abbreviated as XG Boost. The word XG Boost, on the other hand, refers to the engineering goal of pushing the computational resources for boosted tree algorithms to their limits. XG Boost is a software library that may be downloaded and installed on a computer and then accessed through a variety of interfaces.

D. Multinomial NB (MNB)

The Multinomial NB (MNB) algorithm is a possible learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem.

E. Testing

System testing is based on the logical assumption that, if all components of the system are correct, system testing will be useful as a user-oriented vehicle prior to deployment. System testing finds faults, provides a recommendation to the administrator and alters the alteration, as well as checks the output's reliability. Before going live, the system is checked to see if the necessary software and hardware are in place to complete the project. To guarantee that this project is correct, it has passed the following testing methods.

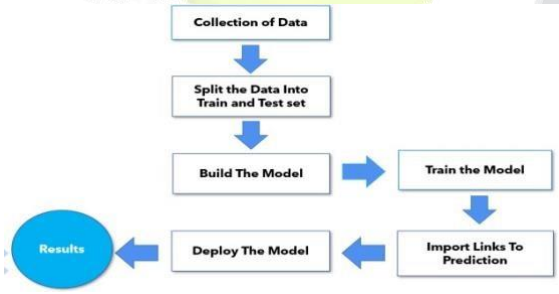

Figure 1. Graphical workflow of proposed models for detection phishing of websites

IV. RESULTS AND DISCUSSION

The current system merely detects phishing websites using multiple machine learning techniques and calculates their accuracy. The best model for detecting phishing websites is generated in the suggested system, and the model is saved and deployed, which takes the URL and predicts whether it is a criminal identity theft website or a real website. When compared to the old approach, the aforementioned statements show that this delivers better accuracy in detecting phishing websites. The accuracy of Logistic Regression is 96.63 percent, and the overall comparison is presented. The overall comparison is given in fig.2.

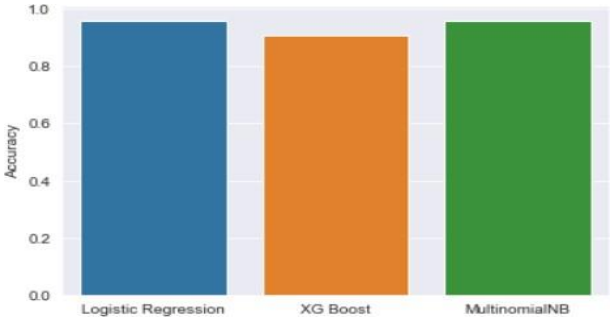

Figure 2. Comparison of the accuracy

V. CONCLUSION

It is remarkable that a good anti-phishing system should be able to predict phishing attacks in a reasonable amount of time. Accepting that having a good anti-phishing gadget available at a reasonable time is also necessary for expanding the scope of phishing site detection. The current system merely detects phishing websites using multiple machine learning techniques and calculates their accuracy.

REFERENCES

[1] Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong, Chengshan Zhang. An Empirical Analysis of Phishing Blacklists. In: Proceedings of the 6th Conference on Email and Anti-Spam (CEAS 2009), Mountain View, California, USA, July 16-17, 2009.
[2] Mahmoud Khonji, Youssef Iraqi, Senior Member, Andrew Jones. Phishing Detection: A Literature Survey. IEEE Communications Surveys and Tutorials, 15(4), 2013, pp.2091-2121. 2013.
[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users‟ Choices Online. ACM Computing Surveys, 50(3), 2017, Article No. 44. 2017.
[4] María M. Moreno-Fernández, Fernando Blanco, Pablo Garaizar, Helena Matute. Fishing for phishers. Improving Internet users' sensitivity to visual deception cues to prevent electronic fraud. Computers in Human Behavior, Vol.69, 2017, pp.421-436.
[5] M. Junger, L. Montoya, F.J. Overink. Priming and warnings are not effective to prevent social engineering attacks. Computers in Human Behavior, Vol.66, 2017, pp.75- 87.2017.
[6] El-Sayed M. El-Alfy. Detection of Phishing Websites Based on Probabilistic Neural Networks and K-Medoids Clustering. The Computer Journal, 60(12), 2017, pp.1745- 1759.
[7] Cheng Huang, Shuang Hao, Luca Invernizzi, Yong Fang, Christopher Kruegel, Giovanni Vigna. Gossip: Automatically Identifying Malicious Domains from Mailing List Discussions. In: Proceedings of the 2017 ACM on Asia Conference on Computer 33 and Communications Security (ASIA CCS 2017), Abu Dhabi, United Arab Emirates, April 2-6, 2017, pp.494-505. 2017.
[8] Frank Vanhoenshoven, Gonzalo Nápoles, Rafael Falcon, Koen Vanhoof, Mario Köppen. Detecting malicious URLs using machine learning techniques. In: Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI 2016), December 6-9, 2016.
[9] Joshua Saxe, Richard Harang, Cody Wild, Hillary Sanders. A Deep Learning Approach to Fast, Format-Agnostic Detection of Malicious Web Content. In: Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops (SPW 2018), San Francisco, CA , USA, August 2, , pp.8- 14. 2018.
[10] Longfei Wu, Xiaojiang Du, Jie Wu. Effective Defense Schemes for Phishing Attacks on Mobile Computing Platforms. IEEE Transactions on Vehicular Technology, 65(8), , pp.6678-6691. 2016.
[11] R. Gowtham, Ilango Krishnamurthi. A comprehensive and efficacious architecture for detecting phishing webpages. Computers & Security, Vol.40, 2014, pp.23-37.
[12] Guang Xiang, Jason I. Hong, Carolyn Penstein Rosé, Lorrie Cranor. CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites. ACM Transactions on Information and System Security, 14(2), , Article No. 21. 2011.
[13] Erzhou Zhu, Chengcheng Ye, Dong Liu, Feng Liu, Futian Wang, Xuejun Li. An Effective Neural Network Phishing Detection Model Based on Optimal Feature Selection. In: Proceedings of the 16th IEEE International Symposium on Parallel and Distributed Processing with Applications Melbourne, Australia, December 11-13, 2018, pp.781-787. (ISPA 2018).
[14] Zuochao Dou,Issa Khalil, AbdallahKhreishah, Ala Al-Fuqaha,MohsenGuizani,“Systematization of Knowledge (SoK): ASystematic Review of Software-Based Web Phishing Detection”,IEEE Communications Surveys & Tutorials, 2017.
[15] Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection andanalysis of drive-by-download attacks and malicious javascriptcode",Proceedings of the 19th International Conference on World WideWeb, pp. 281-290, 2010.
[16] Choon Lin Tan, Kang LengChiew, San Nah Sze, “Phishing WebsiteDetection Using URL-Assisted Brand Name Weighting System”,2014 IEEE International Symposium on Intelligent Signal Processingand Communication Systems (ISPACS) December 1-4, 2014.
[17] R. B. Basnet, A. H. Sung, "Mining web to detect phishing urls",Proceedings of the International Conference on Machine Learning andApplications, vol. 1, pp. 568-573, Dec 2012.
[18] Mohiuddin Ahmed, AbdunNaserMahmood, Jiankun Hu, "A survey ofnetwork anomaly detection techniques", J. Netw. Comput. Appl., vol.60, no. C, pp. 19-31, 2016.
[19] LuongAnh Tuan Nguyen, Ba Lam To, HuuKhuong Nguyen1 andMinh Hoang Nguyen, “A Novel Approach for Phishing DetectionUsing URL-Based Heuristic”, 2014 International Conference onComputing, Management and Telecommunications (ComManTel), IEEE 2014.

[20] S. CarolinJeeva, Elijah Blessing Rajsingh, "Intelligent phishing urldetection using association rule mining", Human-centric Computingand Information Sciences, vol. 6, no. 1, pp. [10] HibaZuhair, Ali Selamat, MazleenaSalleh, "Feature selection forphishing detection: a review of research", International Journal ofIntelligent Systems Technologies and Applications, Vol. 15, No. 2,201610, 2016.
[21] S. Duffner and C. Garcia, "An Online Backpropagation Algorithmwith Validation Error-Based Adaptive Learning Rate," in ArtificialNeural Networks – ICANN 2007, Porto, Portugal, 2007.
[22] R. M. Mohammad, F. Thabtah and L. McCluskey, "Predictingphishing websites based on self-structuring neural network," NeuralComputing and Applications, vol. 25, no. 2, pp. 443-458, 2013-B.
[23] Huang, H., Tan, J. and Liu, L. (2009) ‚Countermeasure techniques fordeceptive phishing attack", International Conference on New Trendsin Information and Service Science (NISS"09), 30 June–02 July, 2009,China, pp.636–641.
[24] Mayuri, A. and Tech, M.(2012) „Phishing detection based on visualsimilarity",International Journal of Scientific and EngineeringResearch (IJSER), Vol. 3, No. 3, March.
[25] Chandrasekaran, Madhusudhanan,Krishnan Narayanan, and ShambhuUpadhyaya. "Phishing email detectionbased on structural properties." NYSCyber Security Conference. 2006.
[26] Wenyin, Liu, et al. "Discoveringphishing target based on semantic linknetwork." Future GenerationComputer Systems 26.3 (2010): 381-388.
[27] Almomani, Ammar, et al. "Evolvingfuzzy neural network for phishingemails detection." Journal ofComputer Science 8.7 (2012): 1099.
[28]. M. Madhuri, K. Yeseswini, and U.VidyaSagar. "Intelligent phishingwebsite detection and preventionsystem by using link guardalgorithm." Int. J. Commun. Netw.Secure (2013): 9-15.
[29] Afroz, Sadia, and Rachel Greenstadt."Phishzoo: Detecting phishingwebsites by looking atthem." Semantic Computing (ICSC),2011 Fifth IEEE InternationalConference on. IEEE, 2011.
[30] Wenyin, Liu, et al. "Discoveringphishing target based on semantic linknetwork." Future GenerationComputer Systems 26.3 (2010): 381-388.
[31] A.Bergholz,J H Chang,G Paass,F Reichartz, &S. Strobel, "ImprovedPhishing Detection using Model-Based Features." CEAS. 2008.