JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

INTELLIGENT DETECTION DESIGNS OF HTML URL PHISHING ATTACKS

[1]V. Sumanth Reddy, [2]G. Vivek, [3]G. Varun, [4]M. Kumar

[1,2,3]UG Scholars , [4]Asst. Professor

[1,2,3,4]Department of Computer Science Engineering (Internet of Things)

Guru Nanak Institutions Technical Campus (Autonomous), Hyderabad, India

Abstract : *- Phishing attacks are a type of cybercrime that has grown in recent years. It is part of social engineering attacks where an attacker deceives users by sending fake messages using social media platforms or emails. Phishing attacks steal users' information or download and install malicious software. They are hard to detect because attackers can design a phishing message that looks legitimate to a user. This message may contain a phishing URL so that even an expert can be a victim. This URL leads the victim to a fake website that steals information, such as login information, payment information, etc. Researchers and engineers work to developmethods to detect phishing attacks without the need for the eyes of experts. Even though many papers discuss HTML and URL-based phishing detection methods, there is no comprehensive survey to discuss these methods. Therefore, this paper comprehensively surveys HTML and URL phishing attacks and detection methods. We review the current state-of-art machine learning models to detect URL-based and hybrid-based phishing attacks in detail. We compare each model based on its data preprocessing, feature extraction, model design, and performance.*

I. INTRODUCTION

Phishing attacks are cybercrime using social engineering to deceive users into stealing theirinformation, such as personal identity, financial information, etc. Masquerading as legitimate sources, attackers can reach victimsby sending fraudulent messages using emails (such as Gmail, Outlook, etc.) or social mediaplatforms (like Twitter, Facebook, etc.). Users become vulnerable if they input their information or download attachment files. In recent years, there has been an increase in social media platform attacks since it is easy for attackers to reach many users from anywhere in the world by posting a single message. According to the Anti-Phishing Working Group (APWG) reports the number of phishing attacks increased by 250000 in onemonth in Jan 2021. In addition, the number of business compromises increased 56% from thelast quarter of 2020 to the first quarter of 2021.The most targeted industries in 2021 are financial institutions, social media, and web emails. Base attackers' primary focus is to stealvictims' financial information or identities by targeting financial industries and social media platforms, respectively. Attackers also might send malicious software that leads to other cyberattacks, such as malware attacks, ransomware attacks, etc. Increasing phishing attacks in recent years and their cybersecurity threats have become an important issue to be solved. Most current organizations rely on using human knowledge to detect these attacks. Nevertheless, phishing attacks are complex for the human eye to identify, even for an expert, due to the similarity between legitimate and fake messages. Therefore, cybersecurity experts pay more attention to the message attachments, such as Uniform Resource Locators (URLs) or email IDs, etc., to recognize phishing messages. Nevertheless, attackers improve their attack techniques by using new methods to design phishing attacks that are hard to detect. For example, they design a phishing URL and webpage that look similar to a benign URL, Therefore, it is essential to determine methods to distinguish a phishing URL from a benign URL. As a result, researchers have proposedseveral solutions in recent years with high accuracy to detect phishing attacks blacklist.

LITERATURE SURVEY

Title: 'Vessel navigation behavior analysis and multiple-trajectory prediction model basedon AIS data

Author: H. Ma, Y. Zuo, and T. Li.

Year: 2022.

Description: With the increasing application and utility of automatic identification systems (AISs), large volumes of AIS data are collected to record vessel navigation. In recentyears, the prediction of vessel trajectories has become one of the hottest research issues. In contrast to existing studies, most researchers have focused on the single-trajectory prediction of vessels. This article proposes a multiple-trajectory prediction model and makes two main contributions. First, we propose a novel method of trajectory feature representation that uses a hierarchical clustering algorithm to analyze and extract thevessel navigation behavior for multiple trajectories. Compared with the classic methods, e.g., Douglas–Peucker (DP) and least-squares cubic spline curve approximation (LCSCA) algorithms, the mean loss of trajectory features extracted by our method is approximately 0.005, and it is reduced by 50% and 30% compared to the DP and LCSCA algorithms, respectively. Second, we design anintegrated model for simultaneous prediction of multiple trajectories using the proposed features and employ the long short-term memory (LSTM)-based neural network and recurrent neural network (RNN) to pursue thistime series task. Furthermore, the comparativeexperiments prove that the mean value and standard deviation of root mean squared error (RMSE) using the LSTM are 4% and 14% lower than those using the RNN, respectively.

Title: Phishing email detection using naturallanguage processing techniques: A literaturesurvey.

Author: S. Salloum, T. Gaber, S. Vadera, and K. Shaalan.

Year: 2021.

Description: Phishing is the most prevalent method of cybercrime that convinces people toprovide sensitive information; for instance, traditional machine learning and DeepLearning (DL). Account IDs, passwords, and bank details. Emails, instant messages, and phone calls arewidely used to launch such cyber-attacks. Despite constant updating of the methods of avoiding such cyber-attacks, the ultimate outcome is currently inadequate. On the other hand, phishing emails have increased exponentially in recent years, which suggests aneed for more effective and advanced methodsto counter them. Numerous methods have beenestablished to filter phishing emails, but the problem still needs a complete solution. To thebest of our knowledge, this is the first survey that focuses on using Natural Language Processing (NLP) and Machine Learning (ML)techniques to detect phishing emails. This study provides an analysis of the numerous state-of-the-art NLP strategies currently in use to identify phishing emails at various stages ofthe attack, with an emphasis on ML strategies. These approaches are subjected to a comparative assessment and analysis. This gives a sense of the problem, its immediate solution space, and the expected future research directions.

Title: Machine Learning Techniques for detection of website phishing: A review forpromises and challenges**.**

Author: A. Odeh, I. Keshta, and E.Abdelfattah.

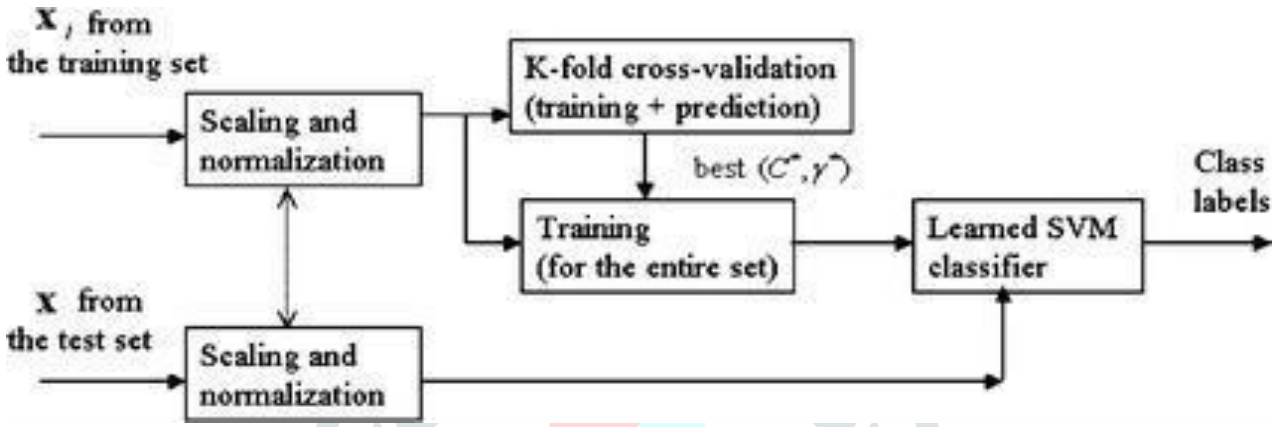
Year: 2021.

Description: — Websites phishing is a cyber- attack that targets online users to steal their sensitive information including login credentials and banking details. Attackers fool the users by presenting the masked webpage aslegitimate or trustworthy to retrieve their essential data. Several solutions to phishing websites attacks have been proposed such as heuristics, blacklist or whitelist, and Machine Learning (ML) based techniques. This paper presents the state of art techniques for phishingwebsite detection using the ML techniques. This research identifies solutions to the website's phishing problem based on the MLtechniques. The majority of the examined approaches are focused on traditional ML techniques. Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Ada Boosting are the powerful ML techniques examined in the literature. This survey paper also identifies deep learning-based techniques with better performance for detecting phishing websites compared to the conventional ML techniques. Challenges to ML techniques identified in this work include overfitting, low accuracy, and ML techniques' ineffectiveness in case of unavailability of enough training data. This research suggests that Internet users should know about phishingto avoid cyber-attacks. This paper also points out the proposal for an automated solution to phishing websites.

II. PROPOSED SYSTEM

Innovative approaches are continually being proposed to fortify phishing detection capabilities. Hybrid models that amalgamate machine learning and featureengineering techniques exhibit promise.For instance, combining lexical and host-based features with machine learning algorithms yields a morecomprehensive understanding of phishing characteristics. Feature-rich models, integrating lexical, content, and host-based features, aim to capture a broader spectrum of malicious behaviors. Furthermore, advancements innatural language processing techniques, and ML models, present opportunities for improved detection by deciphering nuanced patterns in phishing messages. These proposed algorithms leveragediverse strategies to bolster accuracy and resilience against sophisticated phishing strategies.

III. SYSTEM ARCHITECTURE



IV. HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It should be what the system does and not how it should be implemented.

- PROCESSOR : DUAL CORE 2 DUOS.
- RAM : 4GB DDR RAM
- HARD DISK : 250 GB

V. SOFTWARE REQUIREMENTS

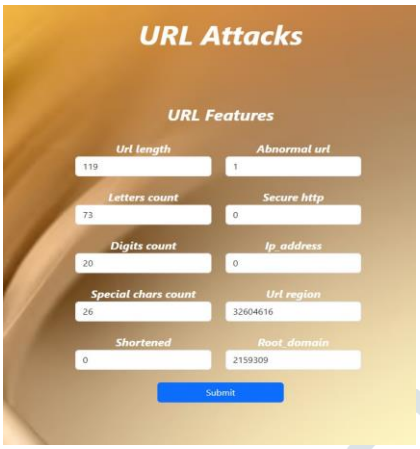
The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the teams and tracking the team's progress throughout the development activity.

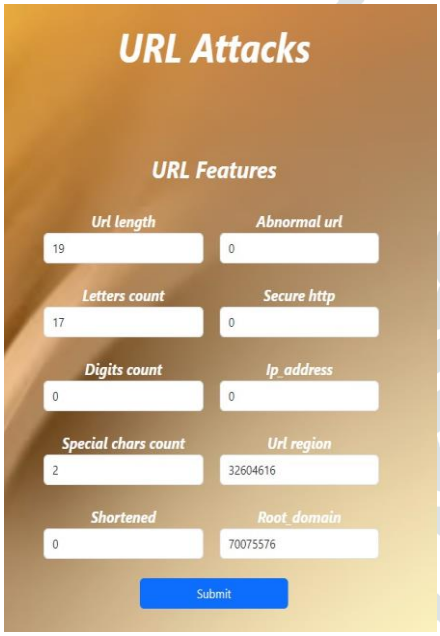
- Operating System : Windows 7/8/10
- Platform : Spyder3
- Programming Language : Python
- Front End : Spyder3

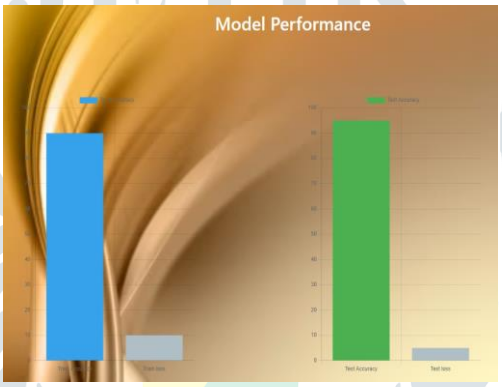
VI. FUTURE ENCHANCEMENT

Feature extraction is where the model reduces the number of features from the input and uses the most related training features. This step differentiates between DL and machine learning. Machine learning needs a human to extract features from the input data. However, DL focuses on learning from the input and its label to extract the most related features. The most common algorithm to extract features in recent years is CNN. The authors propose using CNN for feature extraction.

VII. SNAPSHOTS







VIII. CONCLUSION

In recent years, DL has become essential for solving cybersecurity problems such as phishing attacks. It is due to its ability to extract features from the input data automated instead of manually. This survey studies state-of-art DL models to detect phishing attacks. Its importance lies in analyzing each DL model on every level, from input data to the model output. The importance of data preprocessing is at the same level as the DL model. The data preprocessing affects the model performance in any task, especially once the model implements an application to detect real-time data. For example, the model must be able to classify any input data even if it was not part of the model's dataset. Therefore, in this paper, we pay more attention to data preprocessing and highlight its weaknesses and strengths. Then, we analyze each DL model's design and highlight its strengths and weaknesses.

IX. REFERENCES

[1] Y. Zhang, Y. Xiao, K. Ghaboosi, J. Zhang, and H. Deng, ''A survey of cyber crimes,'' Secure. Common. Network., vol. 5, no. 4, pp. 422–437, 2012.

[2] APWG Developers. (2021). Phishing Activity Trends Report.

[3] M. Lei, Y. Xiao, S. V. Vrbsky, and C.-C. Li, ''Virtual password using random linear functions for on-line services, ATM machines, and pervasive computing,'' Comput. Commun., vol. 31, no. 18, pp. 4367–4375, Dec. 2008.

[5] P. Burda, L. Allodi, and N. Zannone, ''Don't forget the human: A crowdsourced approach to automate response and containment against spear phishing attacks,'' in Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS PW), Sep. 2020, pp. 471–476.

[6] W. Zhang, Y.-X. Ding, Y. Tang, and B. Zhao, ''Malicious web page detection based on on-line learning algorithm,'' in Proc. Int. Conf. Mach. Learn. Cybern., vol. 4, Jul. 2011, pp. 1914–1919.

[7] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González, ''Classifying phishing URLs using recurrent neural networks,'' in Proc. APWG Symp. Electron. Crime Res. (eCrime), 2017, pp. 1–8.

[8] B. Cui, S. He, X. Yao, and P. Shi, ''Malicious URL detection with feature extraction based on machine learning,'' Int. J. High Perform. Comput. Netw., vol. 12, no. 2, pp. 166–178, 2018.

[9] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang, ''Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism,'' IEEE Access, vol. 7, pp. 56329–56340, 2019.

[10] J. Feng, L. Zou, O. Ye, and J. Han, ''Web2Vec: Phishing webpage detection method based on multidimensional

features driven by deep learning,'' IEEE Access, vol. 8, pp. 221214–221224, 2020.
[11] H. Cheng, J. Liu, T. Xu, B. Ren, J. Mao, and W. Zhang, ''Machine learning based low-rate DDoS attack detection for SDN enabled IoT networks,'' Int. J. Sens. Netw., vol. 34, no. 1, pp. 56–69, 2020.
[12] S. Christin, É. Hervet, and N. Lecomte, ''Applications for deep learning in ecology,'' Methods Ecol. Evol., vol. 10, no. 10, pp. 1632–1644, Oct. 2019.
[13] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, ''PhishAri: Automatic real time phishing detection on Twitter,'' in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.
[14] H. Ma, Y. Zuo, and T. Li, ''Vessel navigation behavior analysis and multiple-trajectory prediction model based on AIS data,'' J. Adv. Transp., vol. 2022, pp. 1–10, Jan. 2022.
[15] J. Fang, B. Li, and M. Gao, ''Collaborative filtering recommendation algorithm based on deep neural network fusion,'' Int. J. Sens. Netw., vol. 34, no. 2, pp. 71–80, 2020.
[16] E. S. Gualberto, R. T. De Sousa, T. P. De Brito Vieira, J. P. C. L. Da Costa, and C. G. Duque, ''The answer is in the text: Multi-stage methods for phishing detection based on feature engineering,'' IEEE Access, vol. 8, pp. 223529–223547, 2020.
[17] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, ''Short-term residential load forecasting based on LSTM recurrent neural network,'' IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 841–851, Jan. 2019.
[18] E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu, ''OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network,'' IEEE Access, vol. 7, pp. 73271–73284, 2019.
[19] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, ''Phishing email detection using natural language processing techniques: A literature survey,'' Proc. Comput. Sci., vol. 189, pp. 19–28, Jan. 2021.
[20] M. Korkmaz, O. K. Sahingoz, and B. Diri, ''Feature selections for the classification of webpages to detect phishing attacks: A survey,'' in Proc. Int. Congr. Hum.-Comput. Interact., Optim. Robotic Appl. (HORA), Jun. 2020, pp. 1–9.