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

A Literature Review on Phishing Website detection method based on deep learning framework using Recurrent Neural Network –GRU Model

Mahesh Bagal,Rutuja Ghaskadbi,Komal Londhe,Mansi Netke

Abstract: Phishing attacks typically rely on social networking techniques applied to email or other electronic communication methods. Some methods include direct messages sent over social networks and SMS text messages. We have implemented a deep learning-based framework as a browser plug-in capable of determining whether there is a phishing risk in real-time when the user visits a web page and gives a warning message. The real-time prediction service combines multiple strategies to improve accuracy, reduce false alarm rates, and reduce calculation time, including whitelist filtering, blacklist interception, and machine learning (ML) prediction. The browser plug-in receives client information, calls the background prediction service, and shows the prediction results to users. It is a deep learning-based framework for detecting phishing URLs. We trained and tested the models using seven custom datasets generated from four existing data sources, and we achieved the highest accuracy with the RNN-GRU model. It has a prototype implementation of the proposed framework as a Chrome browser extension.

***Keywords*:** Phishing detection, machine learning, deep learning, RNN-GRU, web browser extension.

I. INTRODUCTION

Phishing is a cybercrime in which a target or targets are contacted by email, telephone or text message by someone posing as a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords. The information is then used to access important accounts and can result in identity theft and financial loss. Other than email and website phishing, there is also 'vishing' (voice phishing), 'smishing' (SMS Phishing) and several other phishing techniques cybercriminals are constantly coming up with. Pharming is a type of phishing attack that uses DNS cache poisoning to redirect users from a legitimate site to a fraudulent one. This is done in an attempt to trick users into attempting to log in to the fake site with personal credentials. With the rapid development of machine learning, there are more and more applications in the field of cybersecurity and we have proposed a deep learning-based framework to detect phishing links in a real-time web browsing environment .We developed a browser plug-in to receive client information, call the background prediction service, and show the prediction results to users. When the URL of the current tab of the browser is predicted to be a phishing link, the current page will receive an obvious warning prompt. The prediction result is obtained by the core prediction service calling a trained machine learning model.

II. LITERATURE REVIEW

The systematic literature review (SLR) was conducted by searching databases of Google Scholar, Web of Science, PubMed, IEEE Xplore Digital Library, PsycInfo and ScienceDirect using the search terms ("Phishing URL Detection") to identify relevant literature. The search strings were run against the title, keywords, and abstract, depending on the search platforms. The searches were conducted between July 1, 2022, through May 20, 2023. Further inputs were also taken from relevant preprints and technical reports. Previous studies have used similar methods to conduct an SLR. To achieve the objectives of extensively reviewing the most relevant studies and answering the research questions. We conducted the SLR under the guidance published. According to Kitchenham and Charters, a Systematic Literature Review is "a form of secondary study that uses a welldefined methodology to identify, analyse and interpret all available evidence related to a specifc research question in a way that is unbiased and repeatable".

1. **Paper Name:** Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing URL detection

Author Name: S.-J. Bu and S.-B. Cho

Description: A deep autoencoder model to detect zero-day phishing attacks and obtained 97.34 % accuracy. They extracted character-level features from URL strings and executed experiments on three different datasets collected from Phish Storm [2], ISCX-URL-2016 , and Phish Tank . They used receiver-operating characteristic curve analysis and N-fold cross-validation to evaluate the experimental results. Comparing the root mean square error (RMSE) in the reconstruction phase between legitimate URLs and phishing URLs, they found the RMSE increased significantly for the phishing URL.

2. **Paper Name:** Web phishing detection using a deep learning framework.

Author Name: D. N. Atimorathanna, T. S. Ranaweera, R. A. H. Devdunie Pabasara, J. R. Perera, and K. Y. Abeywardena

Description: an anti-phishing protection system, which consists of a web browser extension, an e-mail detection plug-in, filters, and a machine learning based phishing detecting server. The browser extension is used to extract the current URL, capture a screenshot, and store the user's visit history as a profile on the client-side. The server mainly uses the following processes to detect phishing links: (1) using the blacklist and whitelist of third-party services to filter new URLs; (2) using a machine learning model based on 13 features to predict whether the URL is a phishing link; (3) using computer vision technology to detect website logos and comparisons the similarity of screenshots of web pages. The logo detector in the article is used to identify 20 well-known online banks and some commonly used website logos.The authors collected and established their own database for the training of the logo detection model and obtained an accuracy rate of more than 95%. The comparison of the similarity of the two screenshots uses Python's OpenCV library. The experimental results of the URL analyzer showed that the Random Forest classifier achieved the highest accuracy of 96.257%. It is a completed online real-time detection system for phishing, combining multiple methods to protect users from being attacked effectively. However, there is still room for improvement in the machine learning model's performance, and the number of logos that the logo classifier can detect is too small.

3. **Paper Name:** Intelligent phishing detection scheme using deep learning algorithms

Author Name: M. A. Adebowale, K. T. Lwin, and M. A. Hossain

Description: It is an combined the convolutional neural network (CNN) and long short-term memory (LSTM) algorithm to classify phishing websites. The hybrid classifier obtained an accuracy of 93.28% and an average computational time of 25s by using image, frame and text features. They collected URLs from Phish Tank and Common Crawl and extracted image features from URLs. The image features are used to feed the offline CNN model, and the text features are contributed to the LSTM classifier. The innovative point of this solution is to combine the characteristics of pictures and text. However, from the experimental results, there is still room for improvement in the accuracy rate, and the calculation time is too long to meet the requirements of realtime prediction products.

4. **Paper Name:** Web phishing detection using a deep learning framework

Author Name: Yi P, Guan Y, Zou F, Yao Y, Wang W and Zhu

Description: a deep learning framework with two types of feature sets, namely original and interaction features. The original features are extracted from the URL analysis, i.e., presence of special characters (@, _, Unicode), count of dots and age of the domain. The interaction features are extracted from the source code of the website, i.e. in-degree, out-degree, frequency of accessing URL and cookie absence. Deep Belief Network (DBN) is applied to the extracted features and achieved an accuracy of 90% true positive rate and 0.6% false positive rate.

5. **Paper Name:**An efficient link-based phishing detection tool

Author Name: .O. Abiodun, A. S. Sodiya, and S. O. Kareem

Description: An website developed to verify a link is a phishing URL or not. The detector was implemented by JAVA programming language and a library named JSoup HTML Parser (JHP). This solution is mainly divided into three stages. The first is to use JSoup to parse the DOM structure of the website to be detected. The second is to analyze the number of link tag <a> from the DOM structure and analyze the attribute ‘‘href’’ value. The attribute value is classified as an empty link, external links and internal links. Third, the link calculator figured out an indicator, which has a value between 0 and 1. When the value exceeds 0.8, the URL to be verified is considered a phishing link. Since no machine learning model is introduced, there is no training process. In the experiment, the authors used 300 URLs to test the performance of the link calculator. The testing results showed they achieved 99.97% accuracy and a 0.03 false-negative rate. They will need to use a larger test data set to verify this solution in the future. From the analysis, it is a misjudgment to judge the phishing risk by analyzing the characteristics of the link tag from the website source code alone, and it is easy for attackers to use this rule to circumvent these rules.

6. **Paper Name:** Development of anti-phishing browser based on random forest and rule of extraction framework

Author Name: M. G. Hr, M. V. Adithya, and S. Vinay

Description: A web browser architecture with an intelligent engine for phishing websites detection named EPDB is presented. Compared to the traditional web browser architectures, the EPDB has a brilliant engine-integrated machine learning model for detection in a real-time environment. They used the UCI dataset to train machine learning models. In the predictive process, the rule of extraction framework is applied, which could extract 30 features of a website. The experimental results showed the Random Forest classifier obtained the highest accuracy of 99.36%. Although the accuracy of the experimental data is very high, this solution also has some limitations and challenges. First, developing a browser is a highly complex task. Some functions of the browser need to be compatible with mature browser functions before they can be promoted to users. In addition, the data set for training the model is single, and the robustness of the model needs to be verified again. Finally, the rule-based feature extraction framework relies on third-party services.

7. **Paper Name:** Efficient deep learning techniques for the detection of phishing websites

Author Name: M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour

Description: An deep learning models for detecting phishing websites only using ten features extracted from HTML and a third-party service. They compared three deep learning models and calculated 18 features' weights. The experimental results demonstrated that the Long Short Term Memory (LSTM) model achieved the highest accuracy of 99.57%. However, they only used one published dataset with 3526 instances. The dataset is obviously too small for deep learning training. The high accuracy rate in the experimental results may be due to the uneven distribution and poor diversity of the test data.

III. CONCLUSION

A deep learning-based framework to detect phishing links in a real-time web browsing environment. An browser plug-in is developed to receive client information, call the background prediction service, and show the prediction results to users. When the URL of the current tab of the browser is predicted to be a phishing link, the current page will receive an obvious warning prompt. The prediction result is obtained by the core prediction service calling a trained machine learning model. It is concluded from the experimental results that the RNN-GRU model obtains the highest accuracy rate of 99.18%.The real-time prediction service combines multiple strategies to improve accuracy, reduce false alarm rates, and reduce calculation time.In this framework a browser plug-in would be capable of determining whether there is a phishing risk in real-time when the user visits a web page and gives a warning message.

IV. REFERENCES

1. S.-J. Bu and S.-B. Cho, ''Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing URL detection,'' *Electronics*, vol. 10, no. 12, p. 1492, Jun. 2021, doi: 10.3390/electronics10121492.

2. D. N. Atimorathanna, T. S. Ranaweera, R. A. H. Devdunie Pabasara, J. R. Perera, and K. Y. Abeywardena, ''NoFish; total anti-phishing protection system,'' in *Proc. 2nd Int. Conf. Advancements Comput. (ICAC)*, Dec. 2020, pp. 470–475, doi: 10.1109/ICAC51239.2020.9357145.

3. M. A. Adebowale, K. T. Lwin, and M. A. Hossain, ''Intelligent phishing detection scheme using deep learning algorithms,'' *J. Enterprise Inf. Manage.*, to be published. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JEIM-01-2020-0036/full/html, doi: 10.1108/jeim-01-2020-0036.

4. Yi P, Guan Y, Zou F, Yao Y, Wang W and Zhu T 2018 Web phishing detection using a deep learning framework. Wirel. Commun. Mobile Comput. 2018: Article ID 4678746
5. O. Abiodun, A. S. Sodiya, and S. O. Kareem, ''Linkcalculator—An efficient link-based phishing detection tool,'' *Acta Inf. Malaysia*, vol. 4, no. 2, pp. 37–44, Oct. 2020, doi: 10.26480/aim.02.2020.37.44.
6. M. G. Hr, M. V. Adithya, and S. Vinay, ''Development of anti-phishing browser based on random forest and rule of extraction framework,'' *Cybersecurity*, vol. 3, no. 1, pp. 1–14, Oct. 2020, doi: 10.1186/s42400-02000059-1.
7. .M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour, ''Efficient deep learning techniques for the detection of phishing websites,'' *Sadhan¯a¯*, vol. 45, no. 1, pp. Jun. 2020, doi: 10.1007/s12046-020-013924.

GUIDE:
Prof.Pragati Pandit

AUTHORS PROFILE

Mahesh Bagal,

K.K.Wagh Institute of Engineering Education And Research,Nashik, India.

Studying in the field of Information Technology

Rutuja Ghaskadbi,

K.K.Wagh Institute of Engineering Education And Research,Nashik, India.

Studying in the field of Information Technology

Komal Londhe,

K.K.Wagh Institute of Engineering Education And Research,Nashik, India.

Studying in the field of Information Technology

Mansi Netke,

K.K.Wagh Institute of Engineering Education And Research,Nashik, India.

Studying in the field of Information Technology