Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

# PHISHING WEBSITE DETECTION IN FUTURE EXTRACTIONS USING NATURAL LANGUGE PROCESS

**Dr.SAMMAIAH SEELOTHU[1], Dr.SHANKARNAYAK BHUKYA[2]**

[1] P.G.T,Department of Computer Science, Maulana Azad National Urdu University , Gachibowli , Hyderabad,Telangana,India , sammaiahmanuu@gmail.com.

[2] Professor, Department of CSE (Data Science) ,CMR Technical Campus, Hyderabad-501401

Telangana ,India , bsnaik546@gmail.com.

## ABSTRACT:

As a crime of employing technical means to steal sensitive information of users, phishing is currently a critical threat facing the Internet, and losses due to phishing are growing steadily. Feature engineering is important in phishing website detection solutions, but the accuracy of detection critically depends on prior knowledge of features. Moreover, although features extracted from different dimensions are more comprehensive, a drawback is that extracting these features requires a large amount of time. To address these limitations, we propose a multidimensional feature phishing detection approach based on a fast detection method by using deep learning (MFPD). In the first step, character sequence features of the given URL are extracted and used for quick classification by deep learning, and this step does not require third-party assistance or any prior knowledge about phishing. In the second step, we combine URL statistical features, webpage code features, webpage text features and the quick classification result of deep learning into multidimensional features. The approach can reduce the detection time for setting a threshold. Testing on a dataset containing millions of phishing URLs and legitimate URLs, the accuracy reaches 98.99%, and the false positive rate is only 0.59%. By reasonably adjusting the threshold, the experimental results show that the detection efficiency can be improved.

***Key words: DDOS, SVM algorithm, phishing,url,Ml,Nlp.***

## I INTRODUCTION

The Internet has become an indispensable infrastructure that brings great convenience to human society. However, the Internet is also characterized by some inevitable security problems, such as phishing, malicious software, and privacy disclosure,

which have already brought serious threats to the economy of users. The APWG (Anti-Phishing Working Group) defines phishing as a criminal mechanism employing both social engineering and technical subterfuge to steal personal identity data and financial account credentials of consumers [1]. Phishing is a very popular method used in network attacks and leads to privacy leaks, identity theft and property damage. According to statistics from the Kaspersky Lab, in 2017, 29.4% of user computers were subjected to at least one Malware-class web attack over the year and 199 455 606 unique URLs were recognized as malicious by web antivirus components [2]. In addition, the share of financial phishing increased from 47.5% to almost 54% of all phishing detections in 2017 [2]. Phishing has become one of the biggest security threats in the Internet. The spread of phishing is no longer limited to traditional modalities such as e-mail, SMS, and pop-ups. Though the prosperity of the mobile Internet and social networks have brought convenience to users, they have also been employed to spread phishing, such as QR code phishing, spear phishing and spoof mobile applications [3], [4], [5], etc. In addition, many cunning phishing attacks are hosted on websites that have HTTPS and SSL certificates because many users think that HTPPS websites are likely legitimate [1]. Phishing presents a diversified development trend, which poses new detection challenges. While phishers are pernicious and hide, security experts and researchers have dedicated many efforts in terms of phishing website detection. Blacklists and whitelists are widely used in phishing website detection. The current common browsers integrate blacklists and whitelists to protect users from phishing attacks. Google provides a blacklist of malicious websites that is continuously updated. Users can check the security of URL links through Google Safe Browsing APIs [6]. Phishing website detection based on blacklists and whitelists is easy to implement with high running speed and a low false positive rate. However, according to statistics [7], 47%-83% of phishing websites are added to blacklists after 12 hours, and 63% of phishing websites have a lifespan of only 2 hours; thus, the updating of the blacklist is far behind the generation of phishing websites. In addition to blacklist and whitelist, machine learning methods are widely used in phishing website detection

[8], [9]. The reason is that malicious URLs or phishing webpages have some characteristics that can be distinguished from legitimate websites, and machine learning can be effective in this regard for processing. Current mainstream machine learning methods of phishing website detection extract statistical features from the URL and the host [10] or extract relevant features of the webpage, such as the layout, CSS, text [11], [12], and then classify these features. However, these methods only analyze the URL or extract features from a single perspective, which makes it difficult to extract the complete attributes of phishing websites. Moreover, some unreasonable features may reduce the accuracy of detection. The character sequence of the URL is natural, automatically generated feature that avoids the subjectivity of artificially selected features. In addition, it does not require third-party assistance and any prior knowledge about phishing. However, in the process of character sequencing, the difficulty is to effectively extract association and semantic information. To address these problems, we propose a multidimensional feature phishing detection approach based on a fast detection method by using deep learning (MFPD). In the first step, character sequence features of the given URL are extracted and used for quick classification by deep learning. Specially, the CNN (convolutional neural network) is used to extract local correlation features through a convolutional layer. In a URL, each character may be related to nearby characters. Generally speaking, a phishing website is likely to mimic the URL of a legitimate website by changing or adding some characters. This can cause the sequential dependency of the phishing URL to be different from the phishing URL. The LSTM network can effectively learn the sequential dependency from character sequences. Therefore, the LSTM (long short-term memory) network is employed to capture context semantic and dependency features of URL character sequences, and at finally softmax is used to classify the extracted features. We call the first step CNN-LSTM. From a comprehensive perspective, in the second step, we combine URL statistical features, webpage code features, webpage text features and the classification result of deep learning into multidimensional features, which are then classified by XGBoost. Although the

multidimensional feature detection method has higher accuracy, it requires extracting features from different aspects, resulting in longer detection time. In contrast, the method for the URL character sequences only needs to process the URL, and the detection time is short. To balance the contradiction between detection time and accuracy, we improve the output judgment condition of the softmax classifier in the deep learning process by setting a threshold to reduce the detection time. If the result of deep learning is not less than the specified threshold, the detection result is directly output; otherwise, go to the second step of detection. In particular, our key contributions in this work are listed as follows: With the phishing website detection as a two-categoryλ processing model, we formally define the problem of phishing detection and give a specific formal description of the MFPD approach. We build a real dataset by crawling a total of 1 021λ 758 phishing URLs as positive samples from phishtank.com, and a total of 989 021 legitimate URLs as negative samples from dmoztools.net. The process of phishing website detection usingλ MFPD is explained, and an extensive experiment on the dataset we built is conducted. The results show that our proposed approach exhibits good performance in terms of accuracy, false positive rate, and speed. A dynamic category decision algorithm (DCDA) isλ proposed. By revising the output judgment conditions of the softmax classifier in the deep learning process and setting a threshold, the detection time can be reduced. The paper is organized as follows. In Section II, we present related work on phishing website detection. Then, in Section III, we introduce the framework of MFPD. In Section IV, we describe the detailed process of the MFPD, which includes the CNN-LSTM and multidimensional features. The performance of the proposed approach is evaluated in Section V. Finally, in Section VI, we conclude the paper and discuss future work.

## II EXISTING SYSTEM

At present, there are some studies on phishing websitedetection based on deep learning. Selvaganapathy et al. [31] proposed a phishing URL detection algorithm using stacked restricted Boltzmann machine for feature selection and deep neural networks as classifiers. Then, multiple detections were constructed using IBK-kNN, Binary Relevance, and Label Powerset with SVM. This model improves the

accuracy of detection by combining the recognition results of multiple classifiers. Bahnsen et al. [32] extracted the syntax and statistical characteristics of the URL, and then classified the character sequence of the URL using LSTM. By comparing with RF, experiments showed that LSTM was better than RF. Based on the above analysis, we regard the URL strings as URL character sequences, which are natural features that do not require prior knowledge about phishing. In the processing of URL character sequences, we refer the idea of the literature [33] to treat the URL as a sequence of text string and quantize the URL at the character level. Therefore, we take advantages of CNN to extract the local features of the sequence, and then use take advantages of LSTM to extract

Nor Ashidi Mat Isa[4] is proposed automated technique for pap smear image using K-Means and Modified seed based region growing algorithm (MSBRG).First, K-means clustering is used to find the threshold value and with this MSBRG is applied for edge detection. As per the result it has given better outcome after comparing with different algorithm for the same. Selvamani.K et.al [5], implemented K-means algorithm to segment the Brain image. This Work is done by Segmenting MRI brain tumour with k tissue values. Estimated mean intensity at each location for each tissue types. Performance of the algorithm is tested using different and large patient data. The future and on-going work is segmenting coronary arteries in a sequence of angiographic image while preserving the topology of the vessel structure.

## III PROPOSED SYSTEM

In this section, we describe the phishing website detection method based on machine learning, including traditional methods and deep learning methods.

The phishing website detection based on machine learning is a hotspot of current phishing website detection research. The results of machine learning methods usually depend on the quality of the extracted features. The focus of current research is on how to extract and select more effective features before processing them.

Resources on the Internet are addressed by URLs, which consist of the Hostname and FreeURL. The typical URL structure is shown in Fig. 1.

Zouina et al. [9] proposed a lightweight phishing website detection method that used only six URL features, namely, the URL size, the number of hyphens, the number of dots, the number of numeric characters plus a discrete variable that corresponds to the presence of an IP address in the URL, and finally, the similarity index. The features extracted are completely based on URLs, and because of their low features,the detection speed is fast. However, the amount of experimental data was relatively small.

Le et al. [15] proposed a method of extracting lexical features from URL strings and using AROW (Adaptive Regularization of Weights) to detect phishing websites. This method overcomes the noise of the training data while ensuring detection accuracy.

## IV METHODOLOGY:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

In this section, we first define the formal statement of phishing website detection, then describe the overall framework of the approach MFPD and its formal definition.

### A. PROBLEM STATEMENT

Suppose we are given a set $U$ consists of all URLs $U = \{u \mid u = x, x = url, i = N\}$, $U = n$. Let $C$as a set indicating phishing, $C_p = \{c \mid c = p, p = phishing\}$ , $C_l$ as a set indicating legitimate, $C_l = \{c \mid c = l, l = legitimate\}$ , and $C = C_p$ , $u_i$ is a suspicious URL. Formally, phishing website detection problem can be defined as follows

### B. THE FRAMEWORK OF MFPD

In this paper, we built the framework of the proposed approach, referred to as MFPD. MFPD can be described by the following four definitions.

(Character embedding of $u_i$). Let the fixed length of the URL $u_i$ be $L$; then, $m = 97 = L$ according to Table 1. For $u_i$, the length of the URL character sequence is unified based on the formula (5) $e_i = URLF(u_i)$ and encoded based on Table 1, $g_i = ASC(e_i)$ . Then, regarding $g_i$ as a vector and $g = (g^1, g^2, ..., g^j)^T$ , $g^j$ indicates the $j$-th elements in the vector, $1 = j = m$ . All URLs form a matrix

$G$, $G = G_{m=n} = (g_1, g_2, ..., g_n)$ . Finally, the embedding network is used to reduce the sparsity of $G$. Letting the network weight be $V$, $V = {}^{p=m}$ , the result of character embedding is

### A. EXPERIMENT DATA AND INDICATORS

The data used in this experiment are real-life data collected from the Internet. First, historical data confirmed as phishing from 2014 to 2018 were crawled from the *PhishTank* website, and a total of 1 021 758 URLs were used as positive samples of the phishing. Then, 989 021 URLs were crawled from the open catalogue website *dmoztools.net* [38] as negative samples of the phishing website, which are legitimate URLs. A total of 2 010 779 URLs were used to set up the dataset *DATA*. Because the survival time of the phishing is short, most of the phishing URLs in *DATA* are not accessible, it is impossible to extract the feature of the webpage code and the text features. To solve this problem, we build the dataset *DATA1* by extracting the currently surviving 22 445 URLs as phishing from *DATA* positive samples, and we randomly select 22 390 accessible URLs from *DATA* negative samples. The remaining data in *DATA* are built into the dataset *DATA2*, which is

*DATA*1 *DATA*2 . *DATA1* is used to verify the effectiveness of the multidimensional feature algorithm and DCDA, and *DATA2* is used to verify the effectiveness of the deep learning algorithm CNN-LSTM.

### B.EXPERIMENT ON THE CNN-LSTM

This experiment is performed on *DATA2* with 5-fold cross- validation. Four sets are used as training sets, the remaining set

is used as a test set. First, the parameters of the CNN- LSTM algorithm need to be adjusted. The experiment finds that the average length of legitimate website samples in dataset *DATA* is 34.7, the average length of phishing website samples is 87.3, the average length of all the data is 61.5, and the length of URLs exceeding 96.3% is below 200. When the number of training epochs reaches 20, the accuracy of the test set is nearly stable; thus, in order to reduce the training time and prevent overfitting, we set *epochs*=20.

To verify the effect of the CNN-LSTM algorithm, three classical deep neural networks, CNN, RNN and LSTM, are compared in this experiment. The structure of the CNN- LSTM algorithm is Input->Conv->Maxpool->LSTM->Softmax. For fairness of the experimental comparison, the network structures of CNN-CNN, RNN-RNN and LSTM- LSTM are compared, whose structure are Input->Conv->Maxpool->Conv->GlobalMaxpool->Softmax, Input->RNN1->RNN2->Softmax, Input->LSTM1->LSTM2-> Softmax, respectively.

We perform calculations on a high-performance server with 64G of memory, a E5-2683 v3 CPU, and GTX 1080ti GPUs, ensuring that deep learning models can be iterated quickly in dealing with large data volumes.

### *C. EXPERIMENT ON THE MULTIDIMENSIONAL FEATURE ALGORITHM*

The effect of the multidimensional feature algorithm is verified in this section. After extracting multidimensional features from *DATA1*, the experiment results using four ensemble learning algorithms for classification are shown in Fig. 15 and Fig. 16. It can be seen that the XGBoost algorithm has the highest accuracy and the lowest FPR, FNR and cost compared with AdaBoost, random forest and GBDT; it also has a faster training speed than GBDT. the statistical feature according to the Table 2, compared with CNN-LSTM and the multidimensional features, as shown in Fig. 17. It can be seen that the multidimensional feature algorithm significantly improves the accuracy and reduces FPR, FNR and cost compared with CNN-LSTM and the traditional feature extraction method.

Table 5 illustrates the three metrics of MFPD and other approaches (J. Mao et al [11], CANTINA+ [19], X. Zhang et al. [32]) based on the evaluation value in the papers. In order to facilitate comparison, we calculate the three metricsbased on our experiment results. X. Zhang et.al [32] has highest recall than MFPD, but MFPD achieves the highest precision and F1. Because the detection process of our approach relies on the hybrid features, which are obtained from multiple aspects and have more information than the features from a single aspect, and it utilizes millions of data for training.

### *D. EXPERIMENT ON THE DYNAMIC CATEGORY DECISION ALGORITHM*

In this section, we conducts five-fold

cross validation on *DATA1* to prove the validity of the dynamic category decision algorithm DCDA. The key of DCDA is to find the optimal threshold α so that it can quickly detect phishing websites with high accuracy and low detection cost.

The experiment results are shown in Fig. 18 and Fig. 19. When a threshold of approximately α=355, the detection accuracy and the detection cost tend to be stable, reaching 98.88% and 4.56, which is almost equivalent to the multidimensional feature detection. The most important role of DCDA is real-time detection. Fig. 20 shows that as the threshold increases, the average number of websites that CNN-LSTM is responsible for detecting gradually decreases, and the number of websites that the multidimensional feature detection is responsible for detecting gradually increases. When the threshold is approximately α=355, only 28% of the websites need to undergo the multidimensional feature detection, which greatly reduces the workload.

## CONCLUSION

It is well known that a good phishing website detection approach should have good real-time performance while ensuring good accuracy and a low false positive rate. Our proposed MFPD approach is consistent with this idea. Under the control of a dynamic category decision algorithm, the URL character sequence without phishing prior knowledge ensures the detection speed, and the multidimensional feature detection ensures the detection accuracy. We conduct a series of experiments on a dataset containing millions of phishing and legitimate URLs. From the results, we find that the MFPD approach is effective with high accuracy, low false positive rate and high detection speed. A future development of our approach will consider applying deep learning to feature extraction of webpage code and webpage text. In addition, we plan to implement our approach into a plugin for embedding in a Web browser.

## REFERENCES

[1] Phishing Attack Trends Re-port-1Q 2018. [Online]. Available: https://apwg.org/resources/apwg-reports/, accessed May. 5, 2018.

[2] Kaspersky Security Bulletin:Overall statisticals for 2017. [Online]. Available: https://securelist.com/ksb-overall-statistics-2017/83453/, accessed Jul.12, 2018.

[3] A. Ahmad Y, M. Selvakumar, A. Mohammed, A. Mohammed and A. S. Samer, "TrustQR: A New Technique for the Detection of Phishing Attacks on QR Code," Adv. Sci. Lett., vol. 22, no. 10, pp. 2905-2909, Oct. 2016.

[4] C. C. Inez and F. Baruch, "Setting Priorities in Behavioral Interventions: An Application to

Reducing Phishing Risk," Risk Anal., vol. 38, no. 4, pp. 826-838, Apr. 2018.

[5] G. Diksha and J. A. Kumar, "Mobile phishing attacks and defence mechanisms: State of art and open research challenges," Comput. Secur., vol. 73, pp. 519-544, Mar. 2018.

[6] Google Safe Browsing APIs. [Online]. Available: https://developers.google.com/safe-browsing/v4/, accessed Oct. 1, 2018.

[7] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An Empirical Analysis of Phishing Blacklists," in Proc. 6th Conf. Email Anti-Spam (CEAS'09), Jul. 2009, pp. 59-78.

[8] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," Eurasip J. Inf. Secur., vol. 2016, no. 1, May. 2016.

[9] M. Zouina and B. Outtaj, "A novel lightweight URL phishing detection system using SVM and similarity index," Human-Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.

[10] E. Buber, Ö. Demir and O. K. Sahingoz, "Feature selections for the machine learning based detection of phishing websites," in Proc. IEEE Int. Artif. Intell. Data Process. Symp. (IDAP), Sep. 2017.

[11] J. Mao, J. Bian, W. Tian, S. Zhu, W. Tao, A. Li and Z. Liang, "Detecting Phishing Websites via Aggregation Analysis of Page Layouts," Procedia Comput. Sci., vol. 129, pp. 224-230, Jan. 2018.

[12] J. Mao, W. Tian, P. Li, T. Wei and Z. Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity," IEEE Access, vol.5, no. 99, pp. 17020-17030, Aug. 2017.

[13] J. Cao, D. Dong, B. Mao and T. Wang, "Phishing detection method based on URL features," J. Southeast Univ.-Engl. Ed., vol. 29, no. 2, pp. 134-138, Jun. 2013.

[14] S. C. Jeeva and E. B. Rajsingh, "Phishing URL detection-based feature selection to classifiers," Int. J. E-Learn. Secur. Digit. Forensics, vol. 9, no. 2, pp. 116-131, Jan. 2017.

[15] A. Le, A. Markopoulou and M. Faloutsos, "PhishDef: URL names say it all," in Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM), Sep. 2010, pp. 191-195.

[16] R. Verma and K. Dyer, "On the character of phishing URLs: Accurate and robust statisticalal learning classifiers," in Proc. 5th ACM Conf. Data Appl. Secur. Priv. (ACM CODASPY), Mar. 2015, pp. 111-122.

[17] Y. Li, S. Chu and R. Xiao, "A pharming attack hybrid detection model based on IP addresses and web content," Optik, vol. 126, no. 2, pp. 234-239, Nov. 2014.

[18] G. Xiang G and J. Hong, "A hybrid phish detection approach by identity discovery and keywords retrieval," in Proc. Int. Conf. World Wide Web (WWW 2009), Oct. 2009, pp. 571-580.

[19] G. Xiang, J. Hong, C. P. Rose and L. Cranor, "Cantina+: A feature- rich machine learning framework for detecting phishing web sites," ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, pp. 21, Sep. 2011.

[20] S. Marchal, K. Saari, N. Singh and N. Asokan, "Know your phish: Novel

techniques for detecting phishing sites and their targets," in Proc. IEEE 36th Int. Conf. Distrib. Comput. Syst. (ICDCS), Jun. 2016, pp. 323-333.