A HYBRID PHISHING DETECTION MODEL BASED ON TRANSFORMER CHARACTERBERT FROM URLS

Conference Paper · November 2021

CITATIONS
0

READS
361

1 author:

Muhammad Sanwal
Akdeniz University
4 PUBLICATIONS **1** CITATION

SEE PROFILE

A Hybrid Phishing Detection Model Based on Transformer CharacterBERT from URLs

Muhammad Sanwal [1] and Alper Ozcan [1]

[1] Akdeniz University, Antalya/Turkey, 202151075006@ogr.akdeniz.edu.tr
[1] Akdeniz University, Antalya/Turkey, alperozcan@akdeniz.edu.tr

***Abstract* – The security concerns in the cybersecurity area are increasing and, the fundamental purpose of such crimes is to obtain the benefits by achieving confidential data. Phishing is one of the easiest ways to accomplish the private information of the target users. Such attacks exist triggered by multiple methods such as emails, messages, phone calls, etc. Therefore, many machine learning techniques have been proposed to detect phishing URLs before affecting the target users. Lately, the popularity of deep learning techniques has also gained attention in the cybersecurity area. This paper proposes a hybrid deep learning model for the classification task between legitimate and phishing URLs. The proposed model consists of CharacterBERT and DNN based techniques. The CharacterBERT is a modification of the baseline BERT model that learns features from the characters of the given URL instead of complete words. Moreover, deep neural networks are applied to train the model. The PhishTank dataset is used for evaluation purposes, and the obtained results indicate that the proposed model outperforms the previous baseline models in the literature.**

***Keywords* – Phishing, Machine Learning, Deep Learning, Deep Neural Networks, CharacterBERT**

I. INTRODUCTION

We live in an era in which everyone is surrounded by technology. The rapid pace of development in technology has created many opportunities in the field of information technology. However, it is not distinct that everyone that interacts with technology will use it positively. The advancement in information technology has generated potential risks to the secrecy of the people where hackers can access the private information by different techniques. Phishing is one of those techniques used by hackers to access potential target information such as username, password, and other information by using several means of communication such as email or messaging. This specific technique encourages the potential target to open the malicious link or download malicious applications on the device. It is a general approach in which malware communications are sent to hundreds of potential targets, and out of all targets, a specific set of users become a victim [1, 2]. Generally, the number of victims is small, but their information is precious. There are different variants of phishing, such as spear-phishing, in which malware is sent to specific individuals or companies. Another type is clone phishing, in which an email is sent with slightly changed original email content that looks legitimate to the potential targets [3]. Moreover, whaling is another type in which communications are made with high-profile targets, and it seems like it is originated from legal departments. Attackers also use search engine optimization techniques in which phishing pages are ranked higher than legitimate web pages.

Generally, the phishing targets are banking services, emails portals, and social media platforms. In the banking service, the typical user provides their debit/credit card credentials, and through phishing, the information can be retrieved by the hackers. People post a large amount of content on social media platforms, and that information is generally more valuable. The attacker can easily fetch detailed knowledge of the target user by using the techniques mentioned above [4].

Phishing is also a common security concern in the health department. The health department contains the information of patients, doctors, and hackers who are keen to access such data, and this information can be used for different kinds of frauds. The IBM report generated in 2018 states that Health Care Industry is more affected by the data breach. Approximately 15 million patients' data was breached in a single year, and phishing was one of them [5]. Another report generated by the US Federal Bureau of Internet Crime in 2019 indicates that a financial loss of $3.5 billion was estimated by such crimes [6].

The number of malicious web pages is minimal compared to web browsers' total number of web pages. However, these numbers are increasing rapidly, and researchers are providing multiple

techniques to detect phishing URLs. Recently, many machine learning and deep learning techniques have been proposed for such tasks. Moreover, in recent times semi-supervised machine learning has also been used for the generation of text embeddings. Such practices are based on natural language processing methods such as BERT based on the transformers and offer significant improvements over the previous methods [7]. In this paper, we proposed a CharacterBERT-DNN based model to detect phishing URLs. Our model comprises of two steps 1) It generates word-level contextual representations by using a features-based mechanism on character embedding 2) it benefits from the hybrid model that is based on DNN and Transformers called CharacterBERT. It is the modified version of the general BERT model.

Furthermore, this paper is organized in the following structure. The Literature review is presented in Section 2, and detailed information of previously proposed important models is provided in Section 3. Moreover, the proposed work is discussed in Section 4. Section 5 describes the datasets and the experiments for this study, and finally, Section 6 describes the future work and conclusion.

II. LITERATURE REVIEW

The popularity of machine learning gained the attention of researchers in the field of cybersecurity. Many machines learning conventional models have been proposed for phishing detection. The machine learning models commonly used in phishing detection are Support Vector Machines, Decision Trees, Random Forest, Bayesian Additive Regression, etc. Generally, the prepared datasets are used in machine learning methods, and these algorithms are beneficial to detect phishing attacks [8]. Sahingoz et al. proposed multiple machine learning algorithms to recognize phishing URLs. Throughout the different sets of experiments, this study achieved high accuracy from Random Forest with Natural Language Processing [9].

Deep learning is evolving rapidly, and deep learning models such as Recurrent Neural Networks, Deep Neural Networks, Convolution Neural Networks, and multi-layer feed-forward networks are commonly used in the literature to detect phishing attacks. The benefit of deep learning models over traditional machine learning models is that such models extract features directly from the given data.

For example, le et al. [10] proposed a deep learning model based on CNN on words and characters of the URLs. This model outperformed the previous models, but there is a possibility of failure if the phishing URL is short. Yi et al. proposed a deep learning framework to extract the original features of the URL, and later they applied Deep Belief Networks to achieve better accuracy [11]. Another study [12] used a self-structuring neural network model for the detection of phishing URLs. In this study, a total of 600 legitimate and 800 phishing websites have been used to build the model. The results of this study indicate that it achieves better results than many studies in this area.

The other commonly used strategy to detect phishing URLs is heuristic feature-based models. This technique predicts the malware URLs by using features in the URL itself or the page content. This model generally classifies the page as malicious or benign. But these types of models tend to perform worse if there are no heuristic features in the websites. Finally, visual Similarity-based models are another detection technique that is used to detect clone websites. Generally, malicious websites are the clone of legitimate websites. The user believes that they are using the legitimate website because it uses the same HTML tags and CSS that generate the overall format of the website. So, such methods compare the features of legitimate and malicious websites and make a verdict on either the website is malicious or legitimate [13].

Pranav et al. proposed a transformers-based model URLTran in which comprehensive analyses are made on the phishing URLs [14]. This study compared standard language masked models and pre-trained domain-specific tasks with fined tuned models such as BERT and RoBERTa. This model improved the performance significantly to detect the false-positive phishing URLs. Moreover, this model was fined adjusted with the adversarial samples to maintain the low false-positive ratio FPRs under controlled scenarios.

Katherine et al. proposed a phishing detection model specifically from website URLs in which the URLs were opened in the mobile devices. They applied ANN as a baseline performance model to HTML and URL-based website features, and this model achieved more than 96% accuracy compared to other states of the art studies in the literature. Further, the deep ANN model was tested only on the URL. The model performed poorly, achieving an accuracy of 86%, indicating that only URL features are not enough to detect phishing websites. Inspired by language transformer, this study applied two states of the art transformers, BERT and ELECTRA, to see phishing websites as the transformers are good at

achieving better contextual information from the text used in the URLs. However, they tested standard and custom vocabularies and found that pre-trained models with fine-tuning achieve better performance for mobile devices than other models [15].

Zhiqiang et al. [16] proposed a dynamic convolutional neural network to extract the malicious features efficiently. A new folding layer is added in the original network that replaces the pooling layer of the web with the kmax-pooling layer. Moreover, this study proposed a new embedding method for word embeddings that are based on character embedding. The character embedding has leverage over the word embedding to learn better vector representation of a URL. The experiments indicate that word embeddings based on character embeddings achieve higher accuracy of 96%.

III. METHODOLOGY

In this section, we will discuss some state-of-the-art studies that are present in the literature. The primary purpose of this section is to provide some details of well-known methods in the Natural Language Processing domain.

A. Phishing URL

A phishing URL contains hurtful words and characters that target legitimate websites to obtain confidential data. In the literature, multiple techniques are mentioned in the literature review section. But in the below part of this section, we try to summarize some of the popular methods extensively used in the literature to provide you with a basic understanding of these models.

B. Transformers

Recently, transformers have been popular and efficient in Natural Language Processing tasks, usually for time series data analysis. Transformers work in a bidirectional manner, and these are popular attention models for language modelling. As transformers work on the encoder-decoder mechanism in which text as an input is provided to the encoder and decoder outputs the possible prediction. A typical transformer model consists of six components: Embeddings, Positional Encoding, Multi-Headed Attention Layers, Feed-Forward Layers, Residual Connections, and Masks. The details of these components can be read from this article [19].

C. LSTM

It is a recurrent neural network architecture that also deals with time-series data. The architecture of LSTM consists of multiple cells that are recursive. This ability allows it to remember and store the information from the previous intervals. The LSTM cell states can be modified by the forget gate and input modulation gate. The unnecessary data can be overlooked from the forget gate, and new information is added. The activation functions play an essential role for each gate. Afterward, the decision is made by the model to classify the input into its respective category. Moreover, LSTM has shown promising results in the field of cyber security, such as phishing detection. One of the applications of LSTM combined with deep neural networks is described in [20], where the model distinguishes between malicious or legitimate URLs.

D. BERT

In the case of BERT, only an encoder is required as it reads the entire sentence at once. This model tries to learn the context of the word by its surroundings. Technically, it is trained on the Masked Language Model (MLM) with Next Sentence Prediction (NSP). The training of MLMs is done by masking out the tokens in the given sentence randomly, and then these masked out tokens are replaced by unique tokens called MASK. Afterward, the model tries to predict the context of these unique masked tokens given in a sequence. Then, NSP models distinguish between input sentences, either those are continuous segments or not. In BERT, the input sequence is tokenized, flows through the encoders, and outputs the hidden state. These hidden states are the representation of word embeddings [10].

E. Convolutional Neural Networks

CNN's have shown promising results in image processing, especially in the image classification tasks. It is the most popular deep learning architecture in the literature. The key to its popularity is it automatically detects the features from the input without supervision. It is composed of convolutional layers that are the building block of its architecture. Generally, the input is provided to the network, which is processed through several convolutional and pooling layers and fully connected layers. The output is the predicted class of the given input. The power of CNNs is recognized in cyber security, and many methods such as [10] use CNNs to classify the URLs into malicious or legitimate. For example, in the study of phishing URLs, the word or characters of the URL are fed.

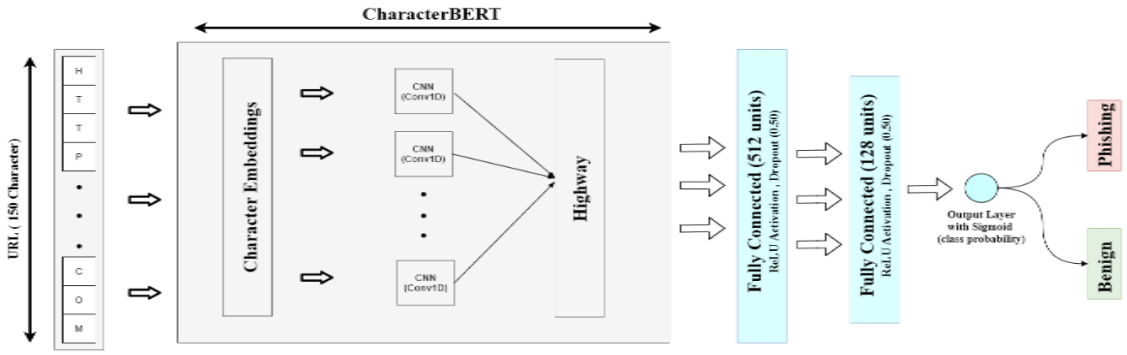

Figure 01: Overview of CharacterBERT model

IV. PROPOSED WORK

In this study, we propose a hybrid model that is the combination of pre-trained transformers and DNN. Precisely for this study, the CharacterBERT transformer is used. This model captures the valuable information and sequential patterns from the URLs of the websites, while most of the other methods require the target websites' content. It uses the sequential string pattern from the URL and classifies them to detect malware or not. The main property of the CharacterBERT model is it learns the representations directly from the URL text and classifies it as legitimate or malicious URLs.

Moreover, it learns non-linear URL embeddings as it is a deep learning framework. In our experiments, this model produced better results than the other baseline models. The below figure illustrates the architecture of the proposed model.

The proposed hybrid model has three layers: 1) CharacterBERT layer, 2) fully connected layers 3) output layer.

A. CharacterBERT layer

It is a new variant of the original BERT model that uses each character to represent the entire word using the Character-CNN module. In contrast, the original model utilizes the whole word for representation in the network. It uses the tokenization technique to represent a single character of the word that is used in [17]. CharacterBERT extracts the features automatically and learns the semantics and patterns from the given URL. Also, we need to define the length of the character in the URL that is being processed in which the longer URLs need trimming, and shorter ones need padding. The padding in the shorter URLs is done by inserting empty characters. The embedding matrix is created with the length of the input sequence of characters. Each character is mapped to its corresponding vector irrespective of the previous or subsequent vectors in the series.

Moreover, the learn embeddings of each character are fed sequentially to the 1-d CNN with filters. Thus, the character embeddings produced earlier are spanned entirely through the CNN filter. Afterward, the max-pooling is done between the character sequence and output layer to create a single representation. Furthermore, the representation from the CNN layers passes through highway layers for non-linearities with residual connections for final embeddings as presented in [17].

B. Fully Connected Layer

To produce a final output, two fully connected layers are defined to receive the input of the CharacterBERT model. Each of these layers is followed by the ReLU activation functions and dropout modules.

C. Output Layer

The last layer only uses the sigmoid function to classify the URLs as either legitimate or malicious.

For novelty, CharacterBERT and DNN algorithms are integrated for phishing detection.

V. EXPERIMENTS AND RESULTS

To compare the proposed work with state-of-the-art studies, we used a PhishTank [18] dataset. This dataset consists of two classes of the URLs legitimate and phishing URLs. As the PhishTank does not

provide free access to the dataset, the Yandex search API is used to find the phishing URLs from the web engines. For this task-specific query, words were constructed by this study [18], and URLs with a very low possibility of phishing were obtained. As the phishing URLs have a concise life, such web pages were among the low-ranked web pages. However, for the testing purpose, the dataset contains 73,575 URLs, of which 36,400 are legitimate, and 37,175 are phishing URLs. To evaluate the performance of the proposed method, we calculated the accuracy, F1-measure value, and AUC as performance metrics. The following formula gives the accuracy:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (1)$$

The F1-measure is the harmonic mean of the precision and recall values. The following formula calculates it.

$$F1 = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) \quad (2)$$

For the experiment of the proposed model, the Google Collab environment is used, and the PyTorch library is preferred for the CharacterBERT and DNN models. We experimented with the datasets on different hyperparameters to obtain the best possible results, and the best values were selected. Moreover, ReLU is used as an activation function and sigmoid as a non-linear function for the output layer. Finally, the dropout rate of 0.5 is selected for all layers, and each fold, 40 epochs are used with a batch size of 128.

The achieved results are illustrated in the following table:

TABLE 1: Comparison of different performance metrics for CharacterBERT-DNN and the two baseline models.

Model	Accuracy (%)	F1-Score (%)	AUC (%)
CNN	95.23	95.12	95.49
LSTM	97.24	97.65	97.30
ELECTRA	91.13	91.40	91.82
BERT	97.10	97.38	97.52
CharacterBERT-DNN	98.41	98.23	98.38

As the results in the above table indicate that CharacterBERT-DNN outperformed both baseline models in this area.

VI. CONCLUSION

In this paper, we proposed a CharacterBERT-DNN based model that classifies the legitimate URLs from Phishing. Most of the existing approaches used word-based embedding techniques that are generally based on the original BERT model. However, as the URL size is small compared to the content of the website, so it is difficult to detect the phishing URLs from the legitimate ones. URL contains different characters that may not create valuable embeddings and contextual meaning for the regular model. The proposed model overcomes this deficiency and uses Character-based embeddings to obtain the maximum possible features from the URL and learns it to classify the phishing URLs. The results from the experiments indicate that our approach is better than the baseline models present in the literature. In the future, we plan to apply ELECTRA based on pre-training text encoders that work as a discriminator rather than the generator in the field of natural language processing.

REFERENCES

[1] CSO Types of phishing attacks and how to identify them. Available: https://www.csoonline.com/article/3234716/phishing/types-of-phishing-attacks-and-how-to-identify-them.html [Accessed 16 Feb 2018].

[2] Priestman, Ward et al. "Phishing in healthcare organisations: threats, mitigation and approaches." BMJ health & care informatics vol. 26,1 (2019): e100031. doi:10.1136/bmjhci-2019-100031.

[3] Abdelhamid M. The role of health concerns in phishing susceptibility: survey design study. J Med Internet Res. 2020;22:e18394.

[4] Maneriker, Pranav & Stokes, Jack & Lazo, Edir & Carutasu, Diana & Tajaddodianfar, Farid & Gururajan, Arun. (2021). URLTran: Improving Phishing URL Detection Using Transformers.

[5] IBM Security. Cost of a Data Breach Report. Traverse City, MI: Ponemon Institute LLC; 2019.
URL: https://www.ibm.com/downloads/cas/ZBZLY7KL [accessed 2020-04-13].

[6] Davis J. Health IT Security. Danvers, MA: intelligent Healthcare Media; 2019. The ten most significant healthcare data breaches of 2019
URL: https://healthitsecurity.com/news/the-10-biggest-healthcare-data-breaches-of-2019-so-far [accessed 2020-04-13].

[7] Ashish Vaswani, Noam Shazier, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. 5998–6008.

[8] Zhang, Y., Hong, J. I. and Cranor, L. F. (2007). Cantina: A content-based approach to detecting phishing websites. 16th Int. World Wide Web Conf. WWW2007, (pp. 639–648). DOI: 10.1145/1242572.1242659.

[9] Sahingoz, O. K., Buber, E., Demir, O. and Diri, B. (2019). Machine learning-based phishing detection from URLs. Expert System Applications. (117), (pp. 345–357). DOI: 10.1016/j.eswa.2018.09.029.

[10] Le, H., Pham, Q., Sahoo, D. and Hoi, S. C. H. (2018). URLNet: Learning a URL representation with deep learning for malicious URL detection. arXiv, no. i.

[11] Yi, P., Guan, Y., Zou, F., Yao, ., Wang, W. and Zhu, T. (2018). Web phishing detection using a deep learning framework. Wireless Communication Mobile Computing. DOI: 10.1155/2018/4678746.

[12] Mohammad, R. M., Thabtah, F. and McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural networks. Neural Computing Application. (25)(2), (pp. 443–458). DOI: 10.1007/s00521-013-1490-z.

[13] Khan, M. F. (2021). Detection of Phishing Websites Using Deep Learning Techniques. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 3880-3892.
[14] Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2021). URLTran: Improving Phishing URL Detection Using Transformers. arXiv preprint arXiv:2106.05256.
[15] Haynes, K., Shirazi, H., & Ray, I. (2021). Lightweight URL-based phishing detection using natural language processing transformers for mobile devices. Procedia Computer Science, 191, 127-134.
[16] Wang, Z., Li, S., Wang, B., Ren, X., & Yang, T. (2020, September). A malicious URL detection model based on a convolutional neural network. In International Symposium on Security and Privacy in Social Networks and Big Data (pp. 34-40). Springer, Singapore.
[17] El Boukkouri Hicham, Ferret Olivier, Lavergne Thomas, Noji Hiroshi, Zweigenbaum Pierre, and Tsujii Junichi. 2020.
[18] Sahingoz, O. K., Buber, E., Demir, Ö., & Diri, B., (2019). Machine learning-based phishing detection from URLs. EXPERT SYSTEMS WITH APPLICATIONS, vol.117, 345-357.
[19] Sabty, C., Islam, M., & Abdennadher, S. (2020). Contextual Embeddings for Arabic-English Code-Switched Data. WANLP.
[20] Ozcan, A., Catal, C., Donmez, E. et al. A hybrid DNN–LSTM model for detecting phishing URLs. Neural Comput & Applic (2021). https://doi.org/10.1007/s00521-021-06401-z.