University of Victoria

Traditional Machine Learning and Deep Neural Network Applications in Phishing

ECE591 Final Project

Dingyang Miao V01028602

Jiaxing Yao V01047304

Lian Duan V01047375

Tom Gan V01032445

Li Zhang V01047097

Lepeng Zhou V01045967

Apr 15, 2024

University of Victoria

Contents

1 Introduction

1.1 Definition and Background of Phishing

Phishing is a sort of fraudulent action that has become more frequent in recent years. Victims are tricked into thinking they were talking with an accredited company. However, they are actually communicating with a fraudster.

These attacks use emails, websites, or messages that appear to be from trustworthy authorities to fool victims disclosing sensitive information or doing risky acts such as clicking on harmful links.

Phishing remains a substantial danger, accounting for 32% of breaches in recent years [1]. It is also the most common type of social engineering attack, accounting for 80% of reported instances [2].

According to the Canadian Anti-Fraud Centre (CAFC) [3], phishing is one of the most common types of fraud. Phishing targets not only individuals, but also a wide range of businesses and government departments. The Canadian Bankers Association (CBA) also regularly warns the public about phishing threats. In the financial sector, phishing attacks attempt to obtain bank account information and credit card details [4]. For example, phishers may send us e-transfer notifications that appear to be real , enticing us to click on the link and enter our banking information.

1.2 Research Purpose

This report aims to explore the role and efficacy of traditional machine learning and deep neural networks in phishing applications. We will examine how different machine learning approaches can be applied to phishing detection and defense processes and what their effectiveness and limitations are. We hope that through this research, a deeper understanding of phishing attacks can be gained, providing insights and suggestions for improving anti-phishing techniques.

2 Traditional Machine Learning

2.1 Introduction to Traditional Machine Learning

Traditional machine learning algorithms perform classification or prediction by learning the regulations of training dataset. Depending on different classification methods, the algorithms use different statistical and mathematical methods. One of the most important process in traditional machine learning is Feature engineering. The process of Feature engineering requires extracting, transforming and selecting features from raw data that are ultimately used for machine learning model training. The quality of feature engineering directly affects the performance and accuracy of the model.

In phishing detection, traditional machine learning algorithms can be used to identify and classify malicious phishing emails, websites, or messages. While the selectable features in feature engineering may include email title, sender address, keywords in the email content, etc.

2.2 Common Machine Learning Models in Phishing

Logistic Regression, K-Nearest Neighbor (KNN), Decision Trees, and Support Vector Machines (SVMs) are four common classification methods for traditional machine learning.

2.2.1 Logistic Regression

Logistic regression is a widely used linear classification algorithm for estimating probabilities between binary values (e.g. 0/1, y/n). It relates the probability of the classification output to the feature variables by using a logistic function. Logistic regression actually solves a probabilistic problem, so it ends up with a probability value and completes the classification by setting a threshold (e.g., 0.5) to convert the probability value to a category value.

2.2.2 K Nearest Neighbor Algorithm (KNN)

In the KNN algorithm, the category of the output is determined by the majority of the categories in its nearest K training samples.The KNN algorithm usually needs to be implemented with consideration of the choice of distances (such as Euclidean distances), as well as the choice of the value of K. One of the main drawbacks of the KNN algorithm is that it is sensitive to outliers.

2.2.3 Decision Trees

A decision tree is a nonparametric supervised learning method that learns simple decision rules to infer the value of a target variable from data features. However, decision trees can be easily overfitted, especially if they are deep, and it may be necessary to set a maximum depth to prevent overfitting.

2.2.4 Support Vector Machines (SVMs)

Support Vector Machines are powerful two-class classification models designed to find an optimal boundary in a dataset that can separate different classes.SVMs work by maximizing the boundary distance between classes and can perform nonlinear classification with kernel functions.SVMs perform well in high-dimensional spaces, and also perform well with respect to external noise and overfitting problems.

2.3 Machine Learning Application in Phishing

With the popularity of Large Language Models (LLMs), criminals are increasingly using LLMs to generate phishing emails that appear legitimate. In a recent study, Greco, F team of researchers attempted to use traditional machine learning models to recognize emails generated either artificially or by LLMs [5]. They selected models such as logistic regression, random forest, SVM, XGBoost, and KNN for comparison. These models were trained and tested on a dataset containing 1000 human-written emails versus 1000 LLM-generated emails.

By analyzing 30 text features, including text consistency, complexity, and other metrics related to AI-generated text, these models were used to distinguish whether an email was LLM-generated or not. The experimental results show that logistic regression models and SVMs excel due to their high accuracy (over 99%) and excellent interpretation capabilities. This finding not only confirms the effectiveness of traditional machine learning techniques in combating advanced AI-generated attacks, but also emphasizes the importance of integrating model interpretation in phishing detection systems in order to enhance users' awareness and ability to defend against phishing attacks.

Through this study, the effectiveness of traditional machine learning models in recognizing phishing emails generated by LLM is well proven. It also makes us realize the importance of incorporating model explanations in the defense strategy to enhance users' defense capability when facing phishing attacks. This result provides a valuable reference for the direction of our future research and enhances our confidence in further developing and improving the anti-phishing model in the next semester.

2.4 Limitation of Machine Learning

The paramount objective in phishing email identification is the accuracy of detection. Conventional machine learning (ML) approaches, like the Random Forest (RF) classifier, have demonstrated their prowess by achieving an average accuracy rate of 99% across various datasets as substantiated in studies [6, 7, 8, 9, 10, 11, 12]. These methodologies typically leverage statistical and lexical features of URLs, such as length and domain age, to differentiate between benign and phishing web pages. The key advantage of ML-based phishing detection is its diminished reliance on blacklists and its capacity to recognize novel malicious URLs [13, 14, 15].

Despite strides in accuracy, ML methods confront significant obstacles: (a) the need for expert manual feature extraction of URLs, sometimes dependent on third-party services for key characteristics; (b) difficulties in managing vast datasets; (c) ineffectiveness against non-standard feature URLs, such as those with short lengths or unique structures; (d) inability to discern semantic patterns, failing to capture all phishing site traits due to a singular evaluative perspective [16].

3 Deep Neural Network (DNN)

3.1 Introduction to DNN

DNN are artificial neural networks composed of multiple layers of artificial neurons, where each layer consists of several neurons that communicate with adjacent layers through weighted connections. DNN are capable of learning complex patterns and representations of input data through the application of nonlinear transformations and optimization of network weights at each layer. This network architecture excels in feature extraction and pattern recognition, and is particularly effective in fields such as image and speech recognition, and natural language processing.

In phishing detection, DNN effectively identify phishing attacks by analyzing and learning features from emails, webpages, or other communication content. The network learns to extract key features from both normal and malicious content to predict the nature of unknown samples.

3.2 Common DNN in Phishing

To surmount the limitations of ML, many researchers [17, 18, 19, 20] have turned to deep learning (DL) techniques, which autonomously extract features without manual intervention. DNN not only match the detection accuracy of traditional ML but also autonomously learn and identify complex features from extensive data sets, making them adept at handling unstructured data like text and images. For instance, they can directly analyze webpage content or structure without manual feature extraction.

Prominent DNN models currently deployed in phishing email detection include:

3.2.1 Autoencoder (AE)

Autoencoder is capable of being trained to reconstruct normal web traffic or user behavior data, these models expose phishing attempts through increased reconstruction error when faced with anomalous input.

3.2.2 Long Short-Term Memory Network (LSTM)

Long Short-Term Memory Network effectively process long sequential data to detect time-dependent features associated with phishing activities within extended user behavior or web request data.

3.2.3 Convolutional Neural Network (CNN)

Convolutional Neural Networ excel in image and text analysis and are used to analyze visual web content, like screenshots, to identify phishing sites masquerading as legitimate entities.

3.2.4 Recurrent Neural Network (RNN)

Recurrent Neural Network's capacity to handle sequential data makes them suitable for analyzing patterns of access to identify phishing websites through the learning and identification of anomalies or suspicious activities in user behavior or web traffic data.

3.3 DNN Applications in Phishing

Cutting-edge studies have increasingly incorporated AE, LSTM, and CNN into phishing detection systems.

A novel approach that marries DNN with AE has been introduced in publication [21]. The strategy involves employing an AE to reduce the dimensionality of the input, subsequently providing a compact output to the DNN for further processing. This combination has proven superior to standalone models, with significant performance gains. By integrating an AE, the accuracy of Artificial Neural Network (ANN) jumped from an estimated 95.6% to around 98%, marking an impressive 3% increase.

In another study [16], Variational Autoencoder (VAE) have been utilized to autonomously derive high-level features from URLs, thus bypassing the traditional reliance on manual feature engineering and third-party feature extraction. This innovation effectively captures intrinsic data from the URLs while also minimizing the dimensions of the input, enabling the detection of malicious URLs that had not been identified before. The streamlined features extracted by the VAE model have reduced training durations for classifiers significantly. The model put forward reached an accuracy rate of 97.45% and demonstrated a rapid response time of 1.9 seconds, surpassing all comparative models in the research.

Further developments have seen the implementation of deep learning models using URL heuristics combined with third-party service-derived features [22]. Employing a variety of deep learning techniques such as CNN, LSTM, and DNN, researchers have been able to effectively discern phishing web entities. The technologies have yielded high accuracy rates—99.57% in LSTM, 99.43% in CNN, and 99.52% in DNN. This method notably utilizes just a fraction of the features provided by third-party services, which minimizes reliance on external feature sets, leads to a reduction in feature volume, enhances robustness, and expedites the phishing detection workflow.

3.4 Limitations of DNN

In recent years, a minority of researchers have utilized neural network-based models for malicious URL detection. Various neural network approaches have been employed to automatically extract inherent features from raw URLs, facilitating unsupervised learning methods. Despite their successes, deep learning-based models face several limitations: (a) URL data must be converted into numerical vectors before being fed into the model. (b) Training the model with raw inputs requires a considerable amount of time. (c) The choice of the neural network model used for feature extraction plays a critical role in the successful categorization of URLs. (d) Depth-based approaches involve a plethora of hyperparameters, and tuning these parameters is a tedious process. (e) The models have slow response times [16].

4 Implementation Plan

Considering the limited size of our dataset, it's expected that SVM will exhibit superior performance. This is largely due to SVM's robust generalization features, which can be optimized through the application of appropriate regularization and kernel techniques.Therefore, we will first attempt to train our phishing detection model using SVM. Additionally, given the scarcity of RNN applications in phishing detection, we also plan to experiment with RNN to assess their effectiveness. The following are the detailed implementation steps:

(1)Data Preprocessing: Cleanse the data of duplicates or missing entries, extract key features, standardize or normalize numerical features, and map vocabularies and pad sequences for textual data (specifically for RNN).

(2)Data Set Division: Split the dataset into 70% for training, 15% for validation, and 15% for testing.

(3)Model Definition: For SVM, select an appropriate kernel function and initial parameters. For RNN, design the architecture and determine the number of layers and neurons.

(4)Model Training: Train the SVM and RNN models with the training set data, using the validation set for parameter optimization.

(5)Performance Evaluation: Compute metrics such as accuracy, precision, recall, and F1 score, and analyze the confusion matrix.

(6)Model Optimization: Adjust model parameters based on evaluation results and iterate to optimize the model.

References

[1] Verizon. (2021). *2021 Data Breach Investigations Report*. *Verizon Business*.

[2] IBM. (2020). *Cost of a Data Breach Report 2020*. *IBM Security*.

[3] Canadian Anti-Fraud Centre. (2020). *Annual Report on Fraud in Canada*. *Public Safety Canada*.

[4] Canadian Bankers Association. (2020). *Cyber Security in Banking*. *Canadian Bankers Association*.

[5] Greco, F., Desolda, G., Esposito, A., & Carelli, A. (2024). *David versus Goliath: Can Machine Learning Detect LLM-Generated Text? A Case Study in the Detection of Phishing Emails*.

[6] Harinahalli Lokesh, G., & BoreGowda, G. (2021). Phishing website detection based on effective machine learning approach. *Journal of Cyber Security Technology*, 5(1), 1–14.

[7] Saleem Raja, A., Vinodini, R., & Kavitha, A. (2021). Lexical features based malicious URL detection using machine learning techniques. *Materials Today: Proceedings*, 47, 163–166.

[8] Gupta, B.B., et al. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications*, 175, 47–57.

[9] Gandotra, E., & Gupta, D. (2021). Improving spoofed website detection using machine learning. *Cybernetics and Systems*, 52(2), 169–190.

[10] Khan, S.A., Khan, W., & Hussain, A. (2020). Phishing Attacks and Websites Classification Using Machine Learning and Multiple Datasets (A Comparative Analysis). In Huang, D.S., & Premaratne, P. (Eds.), *Intelligent Computing Methodologies*. Lecture Notes in Computer Science, vol. 12465. Springer, Cham.

[11] Alam, M. N., Sarma, D., Lima, F. F., Saha, I., Ulfath, R.-E.-., & Hossain, S. (2020). Phishing Attacks Detection using Machine Learning Approach. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1173-1179.

[12] Deshpande, A., Pedamkar, O., Chaudhary, N., & Borde, Dr. S. (2021). Detection of Phishing Websites using Machine Learning. *International Journal of Engineering Research & Technology (IJERT)*, 10(05).

[13] Tang, L., & Mahmoud, Q.H. (2021). A survey of machine learning-based solutions for phishing website detection. *Machine Learning and Knowledge Extraction*, 3(3), 672–694.

[14] Alkawaz, M.H., et al. (2021). A comprehensive survey on identification and analysis of phishing websites based on machine learning methods. In *IEEE 11th Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, pp. 82–87.

[15] da Silva, C.M.R., Feitosa, E.L., & Garcia, V.C. (2020). Heuristic-based strategy for Phishing prediction: a survey of URL-based approach. *Computer Security*, 88, 101613.

[16] Prabakaran, M.K., MeenakshiSundaram, P., & Chandrasekar, A.D. (2023). An enhanced deep learning-based phishing detection mechanism to effectively identify malicious URLs using variational autoencoders. *IET Information Security*, 17(3), 423–440.

[17] Wang, W., et al. (2019). PDRCNN: precise phishing detection with recurrent convolutional neural networks. *Security and Communication Networks*, 2019, 1–15.

[18] Ali, W., & Ahmed, A.A. (2019). Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. *IET Information Security*, 13(6), 659–669.

[19] Yang, L., et al. (2021). An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features. *Expert Systems with Applications*, 165, 113863.

[20] Bu, S.J., & Cho, S.B. (2021). Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing URL detection. *Electronics*, 10(12), 1492.

[21] Gopal, S.B., Poongodi, C., Nanthiya, D., et al. (2023). Autoencoder-Based Architecture for Identification and Mitigating Phishing URL Attack in IoT Using DNN. *Journal of The Institution of Engineers (India): Series B*, 104, 1227–1240.

[22] Somesha, M., Pais, A.R., Rao, R.S., et al. (2020). Efficient deep learning techniques for the detection of phishing websites. *Sādhanā*, 45, 165.