

A Novel Phishing Website Detection Approach Using Machine Learning

Akshay Pawar, Rahul Gaigawale, Suyog Late, Ankit Sharma, Prof. Tanvi Deshmukh

Department of Information Technology, Professor

Sandip Institute of Technology and Research Centre, Nashik

Abstract—Phishing attackers spread phishing links through e-mail, text messages, and social media platforms. They use social engineering skills to trick users into visiting phishing websites and entering crucial personal information. In the end, the stolen personal information is used to defraud the trust of regular websites or financial institutions to obtain illegal benefits. With the development and applications of machine learning technology, many machine learning-based solutions for detecting phishing have been proposed. Some solutions are based on the features extracted by rules, and some of the features need to rely on third-party services, which will cause instability and time-consuming issues in the prediction service. In this project, we propose a machine learning-based framework for detecting phishing websites. We have implemented the 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. In the ML prediction module, we compared multiple machine learning models using several datasets.

Index Terms—Phishing, Personal information, Machine Learning, Malicious links, Phishing domain characteristics, Phishing attacks, legitimate, trust worthy

I. INTRODUCTION

Phishing is a kind of Cybercrime trying to obtain important or confidential information from users which is usually carried out by creating a counterfeit website that mimics a legitimate website. Phishing attacks employ a variety of techniques such as link manipulation, filter evasion, website forgery, covert redirect, and social engineering. The most common approach is to set up a spoofing web page that imitates a legitimate website. These type of attacks were top concerns in the latest 2018 Internet Crime Report, issued by the U.S. Federal Bureau of Investigations Internet Crime Complaint Center (IC3). The statistics gathered by the FBI's IC3 for 2018 showed that internet-based theft, fraud, and exploitation remain pervasive and were responsible for a staggering 2.7 billion in financial losses in 2018. In that year, the IC3 received 20,373 complaints against business email compromise (BEC) and email account compromise (EAC), with losses of more than 1.2 billion. The report notes that the number of these sophisticated attacks have grown increasingly in recent years. Anti-Phishing Working Group (APWG) emphasizes that phishing attacks have grown in recent years, Figure 1 illustrates the total number of phishing sites detected by APWG in the first quarter of 2020 and the last quarter of 2019. This number has a gradual growth raising from 162,155 in the last quarter of 2019 to 165,772

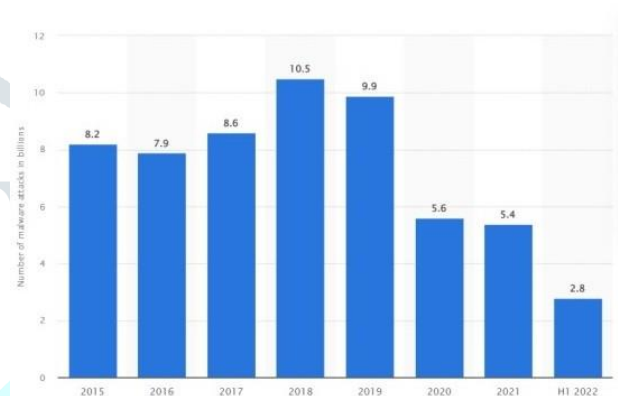


Fig. 1. Graph Design

cases in the first quarter of 2020. Phishing has caused severe damages to many organizations and the global economy, in the fourth quarter of 2019, APWG member OpSec Security found that SaaS and webmail sites remained the most frequent targets of phishing attacks. Phishers continue to harvest credentials from these targets by operating BEC and subsequently gain access to corporate SaaS accounts. Many approaches have been used to filter out phishing websites. Each of these methods is applicable on different stages of attack flow, for example, network-level protection, authentication, client-side tool, user education, server-side filters, and classifiers.

II. EASE OF USE

A. Problem Definition

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. Phishing may be a style of broad extortion that happens once pernicious website act sort of a real one memory that the last word objective to accumulate unstable info, as an example, passwords, account focal points, or MasterCard numbers.

B. Objective

The main objective of this research is to develop a machine learning-based phishing detection system that can help users to check the legitimacy and maliciousness of an URL within a minimum amount of time. To develop a novel approach to detect malicious URL and alert users. To apply ML techniques in the proposed approach in order to analyse the real time URLs and produce effective results. To implement the concept of RNN, which is a familiar ML technique that has the capability to handle huge amount of data

III. LITERATURE SURVEY

Author [1] survey introduced the lifecycle of phishing to clarify the important steps for antiphishing. This paper focuses on the technical methodologies, particularly machine learning-based solutions for phishing website detection. Furthermore, the architecture of machine learning-based resolution shows the general components in the system. The details of each part inspire the development of high-performance phishing detection techniques. We reviewed diverse academic articles and sorted diverse data sources as shown. It is easy to start with published datasets that are standardized based on rules generated by security experts' experience. However, these datasets contain limited instances. Small datasets affect model performance in the training process, particularly for complex structured models such as multi-layer neural networks. In addition, they are relatively old, being collected approximately five years ago.

Author [2] Research proposes meta-heuristic based approach to protect the internet users from the web-phishing. It consists of three phases, the first phase uses a new proposed method for evaluating and ranking the features of URL, HTML and JavaScript code, text, images and domain name of the web page. The second phase extracts the effective subset of the ranked features that achieves the highest classification accuracy of the web-phishing. The third phase constructs the Random forest classifier training by data features of the extracted subset.

Author [3] We started with an overview of phishing attacks and several existing approaches to detect this attack. We further described the limitation of all the existing approaches and formulated our novel approach in phishing detection. Our approach used only nine features based on the lexical properties of URLs and produced an accuracy of 99.57. To clearly extract the features from URLs, we also described the feature extraction algorithms in brief. Later on, we provided the distribution.

IV. SYSTEM ARCHITECTURE

The system consists of major steps preprocessing, feature extraction, and classification. In the testing phase verification is done with pertained sample signatures.

• **Preprocessing** : The motivation behind the pre-processing stage is to make signature standards and prepared for include extraction. The pre-preprocessing stage basically includes

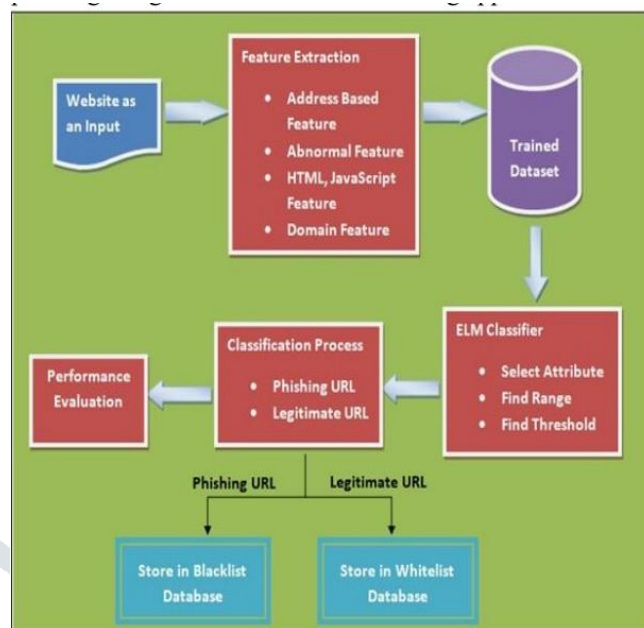


Fig. 2. System Architecture



Fig. 3. Data Flow Diagram

noise, resizing, Binarization, thinning, clutter removal, and normalization.

• **Feature Extraction** : Features extraction is required when input information to an algorithm is excessively huge and repetitive. This excess information is then changed into the brief and fundamental arrangement of features. This technique is called feature extraction.

• **Classification** : Classification is the process where input information is sorted. Another piece of information when contributing to the framework tends to be effectively recognized as having a place with a specific class.

• **Verification** : In this step prepared classifier verify the test signature against a set of test sample signature it has pertained to during the classification stage. If the match is found over a certain threshold, then the signature is considered original else it is considered forged.

V. IMPLEMENTATION

A. Introduction

The methodology involves building a training set. The training set is used for training a machine learning model, i.e., the classifier. Fig 4 shows the diagrammatic representation of the implementation.

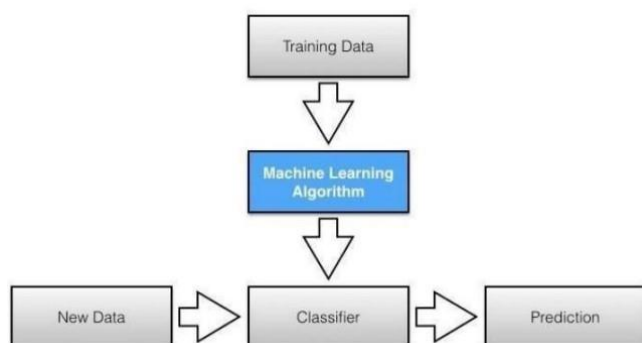


Fig. 4. Implementation

B. Technology Used

Python

In technical terms, Python is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options. Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers.

MACHINE LEARNING

Machine learning provides simplified and efficient methods for data analysis. It has indicated promising outcomes in real time classification problems recently. The key advantage of machine learning is the ability to create flexible models for specific tasks like phishing detection. Since phishing is a classification problem, Machine learning models can be used as a powerful tool. Machine learning models could adapt to changes quickly to identify patterns of fraudulent transactions that help to develop a learning-based identification system. Most of the machine learning models discussed here are classified as supervised machine learning, this is where an algorithm tries to learn a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data. In 2008, developer Wes McKinney started developing

pandas when in need of high performance, flexible tool for analysis of data. Prior to Pandas, Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

C. Flowchart of the system

A flowchart is a diagram that depicts a process, system, or computer algorithm. It is a graphical representation of the steps that are to be performed in a system, it shows the steps in sequential order. It is used in presenting the flow of algorithms and to communicate complex processes in clear, easy-to-understand diagrams. Figure 5 shows the flow of phishing detection systems using the machine learning process.



Fig. 5. Flowchart of the proposed System

Figure 6 shows the phishing detection web interface system. The user inputs a URL link and the website validates the format of the URL and then predicts if the link is phishing or legitimate.

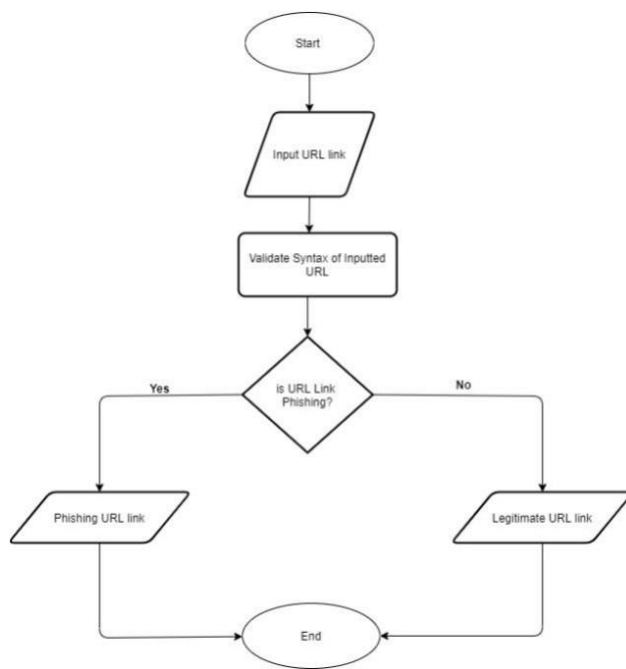


Fig. 6. Flowchart of the web interface

VI. SOFTWARE & HARDWARE REQUIREMENTS

- Windows 7 or higher
- Python 3.6.0 or higher
- Visual Studio Code
- Dataset of Phishing Websites
- Django
- HTML
- FLASK
- 2GB RAM (minimum)
- 100GB HDD (minimum)
- Intel 1.66 GHz Processor Pentium 4 (minimum)
- Internet Connectivity

VII. CONCLUSION

The demonstration of phishing is turning into an advanced danger to this quickly developing universe of innovation. Today, every nation is focusing on cashless exchanges, business online, tickets that are paperless and so on to update with the growing world. Yet phishing is turning into an impediment to this advancement. Individuals are not feeling web is dependable now. It is conceivable to utilize AI to get information and assemble extraordinary information items. A lay person, completely unconscious of how to recognize a security danger shall never invite the danger of making money related exchanges on the web. Phishers are focusing on installment industry and cloud benefits the most. The project means to investigate this region by indicating an utilization instance of recognizing phishing sites utilizing ML. It aimed to

build a phishing detection mechanism using machine learning tools and techniques which is efficient, accurate and cost effective

REFERENCES

- [1] L. Tang and Q. H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 672–694, Aug. 2021, doi:10.3390/make3030034.
- [2] S. Marchal, J. Francois, R. State, and T. Engel, "PhishStorm: Detecting phishing with streaming analytics," *IEEE Trans. Netw. Service Manage.*, vol. 11, no. 4, pp. 458–471, Dec. 2014.
- [3] (Jun. 2021). Phishing Activity Trends Report 1st Quarter 2021. APWG. Accessed: Oct. 20, 2021. [Online]. Available: <https://docs.apwg.org/reports/apwgtrendsreportq12021.pdf>
- [4]] (2020). 2020 Internet Crime Report. [Online]. Available: <https://www.ic3.gov/Media/PDF/AnnualReport/2020IC3Report.pdf>
- [5] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on selfstructuring neural network," *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, Nov. 2013, doi:10.1007/s00521-013-1490-z.
- [6] M. A. El-Rashidy, "A smart model for web phishing detection based on new proposed feature selection technique," *Menoufia J. Electron. Eng. Res.*, vol. 30, no. 1, pp. 97–104, Jan. 2021, doi: 10.21608/mjeer.2021.146286.
- [7] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," *Comput. Commun.*, vol. 175, pp. 47–57, Jul. 2021, doi: 10.1016/j.comcom.2021.04.023.
- [8] E. Gandotra and D. Gupta, "Improving spoofed website detection using machine learning," *Cybern.Syst.*, vol. 52, no. 2, pp. 169–190, Oct. 2020, doi: 10.1080/01969722.2020.1826659.
- [9] W. Wang, F. Zhang, X. Luo, and S. Zhang, "PDRCNN: Precise phishing detection with recurrent convolutional neural networks," *Secur. Commun. Netw.*, vol. 2019, pp. 1–15, Oct. 2019, doi:10.1155/2019/2595794
- [10] M. Sabahno and F. Safara, "ISHO: Improved spotted hyena optimization algorithm for phishing website detection," *Multimedia Tools Appl.*, Mar. 2021, doi: 10.1007/s11042-021-10678-6