IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, May 2022

IJARSCT

Impact Factor: 6.252

Detecting Phishing Websites using Machine Learning

Ninand More[1], Dhruv Solanki[2], Devang Solanki[3], Shanil Jain[4], Praajwal Kadu[5]

Students, Department of Information Technology[2,3,4,5]
Assistant Professor, Department of Information Technology[1]
D. Y. Patil College of Engineering, Pune, Maharashtra, India

Abstract: *The criminals, who want to obtain sensitive data, first create unauthorized replicas of a real website and e-mail. The e-mail will be created using logos and slogans of a legitimate company. The nature of website creation is one of the reasons that the Internet has grown so rapidly as a communication medium. Phisher then send the "spoofed" e-mails to as many people as possible in an attempt to lure them into the scheme. When these e-mails are opened or when a link in the mail is clicked, the consumers are redirected to a spoofed website, appearing to be from the legitimate entity. We discuss the methods used for detection of phishing Web sites based on url importance properties.*

Keywords: Machine Learning, Phishing Website Detection, Phishing, URL Detection

I. Introduction

The internet is growing very fast and everyone is using it because it can afford all needs online as well as providing a wide range of services. As the use of the internet rapidly increase the requirement of providing confidential and critical data transmissions raises. This has revealed it to a wide range of security attacks, Therefore, it is important to implement web applications security techniques while developing an online web application. Besides these attacks, phishing is one of the most common social engineering attacks that is used by attackers to steal critical and sensitive information by disguising themselves as trustworthy organizations, phishing has compromised millions of users' data. Log files are files that keep a registry of events and activities, therefore they are used to save all the logs of the website in order to keep track of every user attempts to log in. The paper is structured as follows: section II discusses problem description, section III describes common Web Application Attacks, and their mitigations are discussed in section IV. Section V illustrates access logs, and in section VI phishing links detection is discussed using machine learning in section VII and deep learning in section VIII. The results are discussed in section IX. Finally, the conclusion and future work are provided in section X.

This study was motivated by the multiple millions of dollars that have been lost due to fraudsters operating fake versions of data collection websites and the need for a safer internet experience as we progress in the internet and communication age.

1.1 Motivation of the Project

The principal motivation behind the task is to recognize the phony or phishing sites who are attempting to gain admittance to touchy information or by making the phony sites and attempting to get access to client individual certification. We are utilizing Al calculation to defend the touchy information and to recognize the phishing sites who are attempting to obtain entrance on delicate information.

II. Description Of The Problem

2.1 Problem Statement

To overcome this problem we are using some of the machine learning algorithms in which will help us to identify the phishing websites based on the features present in the algorithm. By using these algorithms we can be able to keep the user's personal credentials or sensitive data safe from intruders.

DOI: 10.48175/568

2.2 Goals and Objectives

- Use of features extracted from websites that explain characteristics of a website for phishing detection.
- Classification of websites based on such features, using Extreme Learning Machines(ELM) which is an advanced neural network leveraging generalization capabilities given by randomization of weights.
- Developing a phishing detection system
- Creating a reporting platform for other users of the platform to report fake websites in order to build the knowledge base.
- Studying previous work on the proposed topic and looking for ways to improve them.
- Optimizing the system.
- Implementing security standards with the system.
- Creating a system which can also give suggestions to guest users

2.3 Statement of Scope

The main purpose of the scope definition is to clearly describe the boundaries of your project. Scope defines the sides of the box and separates what is relevant to your project from that which is irrelevant.

2.4 Major Constraints

Major constraint is how to get an unbounded real time stream of structured, unstructured, semi structured data that is audio, video, textual. Connective- it's of that big data with different modules framework. To perform analysis and processing on that real-time data. Relevant mathematics associated with the Project

- Input: URL.
- Output: By using a machine learning algorithm it shows reports and dashboards that the website is legitimate or not.
- Success: this will show that which algorithm has maximum efficiency among the algorithm which we are nusing
- Failure: When the internet connection gets loose it will get disconnected.

III. Methodology

3.1 Machine Learning

Writing a review is the most critical advance in the programming improvement process. Before building up the instrument it is important to decide the time factor, economy and friends quality. When these things are fulfilled, at that point following stages is to figure out which working framework and dialect can be utilized for building up the instrument. When the developers begin fabricating the instrument the software engineers require part of outside help. This help can be gotten from senior software engineers, from book or from sites. Before building the framework the above thought are considered for building up the proposed framework.

AI (ML) is a class of calculation that enables programming applications to turn out to be progressively precise in anticipating results without being expressly customized. The fundamental reason of AI is to assemble calculations that can get input information and utilize factual examination to foresee a yield while refreshing yields as new information winds up accessible. The procedures engaged with AI are like that of information mining and prescient displaying. Both require scanning through information to search for examples and modifying program activities as needs be. Numerous individuals know about AI from shopping on the web and being served advertisements identified with their buy. This happens on the grounds that suggestion motors use AI to customize online promotion conveyance in practically continuous. Past customized advertising, other regular AI use cases incorporate misrepresentation location, spam separating, arrange security risk identification, prescient support and building news sources.

3.2 Applications

For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

DOI: 10.48175/568

IV. Literature Review

In this section, we review some of the recent existing works that applied some feature selection methods with machine learning techniques to enhance the detection of phishing websites. Generally, the feature selection methods utilized in detecting phishing websites can be categorized into four categories: frequency analysis-based feature selection, filter-based feature selection, wrapper-based feature selection, and evolutionary algorithm-based feature selection. Many research works have utilized frequency analysis based feature selection to find significant features to improve the performance of intelligent methods in recognizing the legitimate from phishing websites. In [26], the authors assessed many websites' features using a software tool to compute each feature frequency, which represents the feature importance. In [17], seventeen significant features were identified based on frequency analysis. The selected features were used to train self-structuring neural networks in order to distinguish between phishing websites and legitimate ones. In a similar way to [17], [24] analyzed the frequency of websites' features to select the most popular features of websites. Then, rule-based data mining classification models were trained based on the selected website features to recognize the new phishing websites. Find function was exploited by [25] to investigate the most substantial features that exist frequently in numerous websites. Neuro-Fuzzy was then trained with the best five features to detect the phishing websites through an online transaction.

Alternatively, several recent existing works demonstrated that the filter-based feature selection techniques enhanced noticeably the performance of intelligent phishing detection approaches. In[7], the authors exploited both frequency analysis and Chi-Square to select a minimal set of relevant website features from the original features. Based on the selected web site's features, a MCAC (Multi-label Classifier based Associative Classification) model was trained and developed to distinguish the phishing websites from legitimate ones. Information Gain (IG), Chi-square, and Correlation Feature Set were employed by [30] to find the most significant website's features in order to enhance the detection accuracy of phishing websites for some rule-based classification machine learning algorithms: C4.5, RIPPER, and PART. In [8], the authors suggested using the IG, Chi-square, and Correlation Features Set (CFS) to reduce the data dimensionality and select the minimal set of important features. Then, four rule-based classification algorithms (OneRule, JRip,Part, and J48) were trained after applying feature selection methods in order to maximize the detection rate of phishing emails.

The results in the studies mentioned earlier showed that some machine learning algorithms based on filter-based feature selection achieved better detection accuracy of phishing websites and emails. However, other machine learning algorithms that applied the filter methods may suffer from relatively poor performance since the filter-based feature selection methods utilize statistical measures to rate each feature independently of a specific machine learning algorithm.

The wrapper feature selection method coupled with the best-first forward searching method was also applied in phishing email classification by [31] and then compared against IG, Relief-F, and CFS. In [31], the authors demonstrated that the wrapper feature selection method out performed IG, Relief-F, and CFS. To identify phishing websites accurately, the most significant websites' features were selected in [10] by using the wrapper-based feature selection. Accordingly, the training dataset with the selected features was used to train RBFN, SVM, NB, C4.5, kNN, and RF. Results indicated that RBFN, SVM, NB, C4.5, kNN, and RF with the considering wrapper-based feature selection accomplished better detection accuracy compared to these machine learning classifiers based on IG and PCA (Principal Component Analysis). However, the wrapper-based feature selection depends on the machine learning algorithm itself and may be computationally expensive.

Recently, genetic algorithm-based feature selection was used in [29] to find more relevant features in order to enhance the detection accuracy of the machine learning model in phishing websites detection. Although the machine learning techniques with applying GA-based feature selection performed better detection accuracy compared to the same machine learning techniques with other feature selection methods, GA-based feature selection required a longer time for some machine learning algorithms.

V. System Design And Flow

The proposed methodology which imports dataset of phishing and legitimate URLs from the database and the imported data is pre-processed. Detecting phishing websites is performed based on four categories of URL features: domain based, address based, abnormal based and HTML, JavaScript features. These URL features are extracted with processed data and values for each URL attribute are generated. The analysis of URL is performed by machine learning technique which computes range value and the threshold value for URL attributes. Then it is classified into phishing and legitimate URLs. The attribute values are computed using feature extraction of phishing websites and it is used to identify the range value and

DOI: 10.48175/568

threshold value. The value for each phishing attribute is ranging from f-1, 0, 1g; these values are defined as low, medium and high according to phishing website features. The classification of phishing and legitimate websites is based on the values of attributes extracted using four types of phishing categories and a machine learning approach.

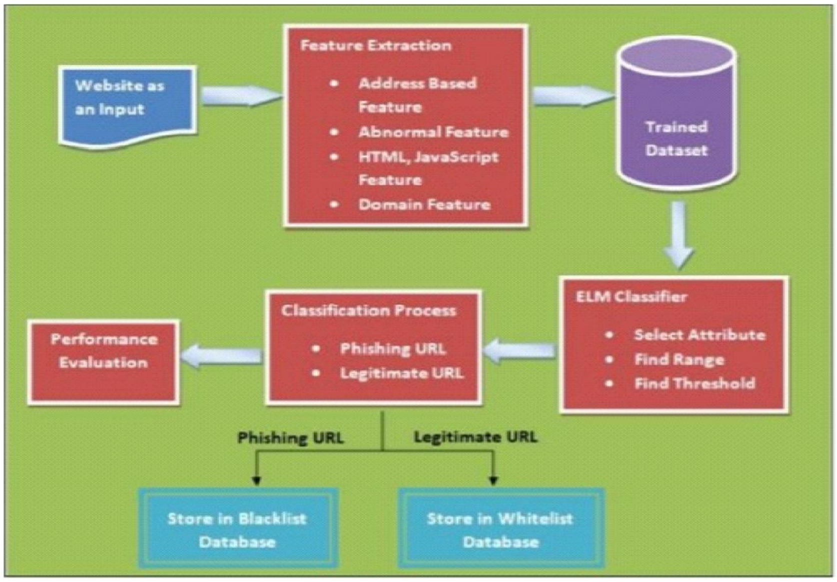


Figure: System Architecture

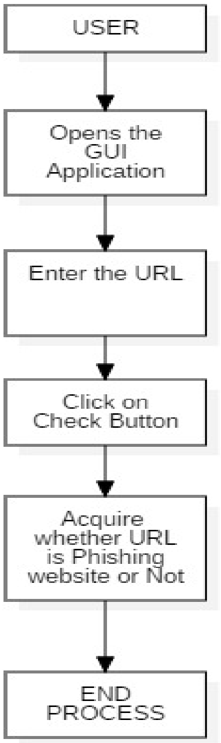
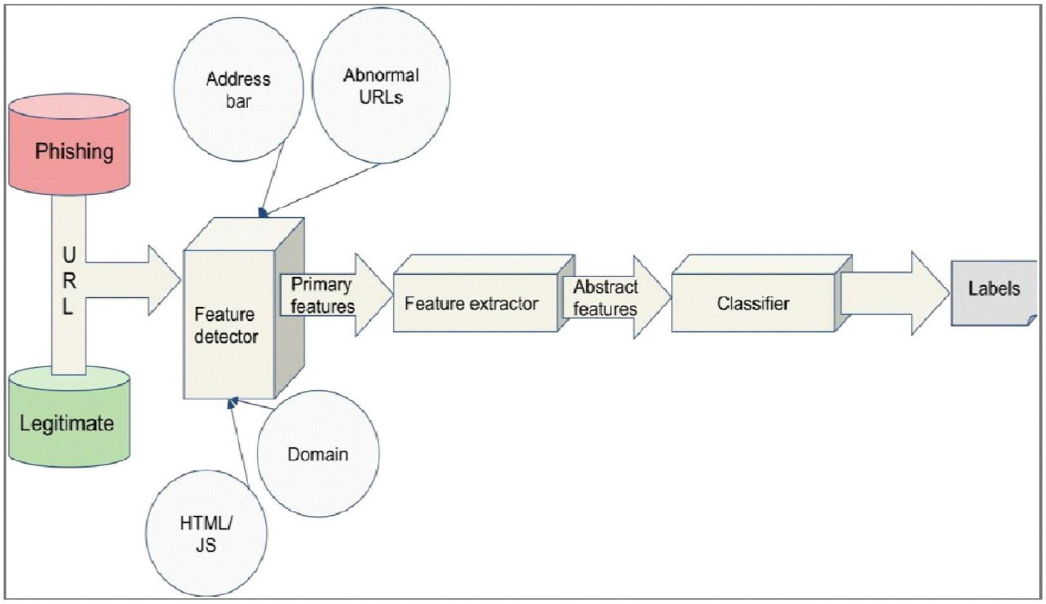

USER
Opens the GUI Application
Enter the URL
Click on Check Button
Acquire whether URL is Phishing website or Not
END PROCESS

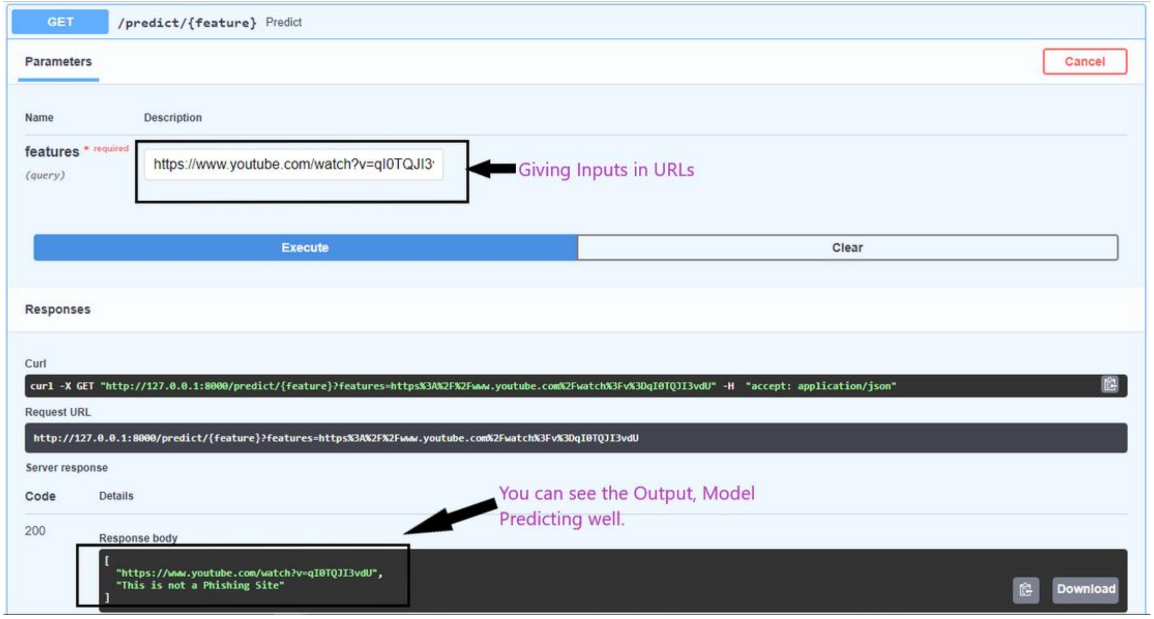
Figure: Data Flow Diagram

DOI: 10.48175/568

VI. PROJECT IMPLEMENTATION



The plan contains an overview of the system, a brief description of the major tasks involved in the implementation, the overall resources needed to support the implementation effort and any site-specific implementation requirements.



https://www.youtube.com/watch?v=qI0TQJI3

Giving Inputs in URLs

Curl

```
curl -X GET "http://127.0.0.1:8000/predict/{feature}?features=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DqI0TQJI3vdU" -H  "accept: application/json"
```

Request URL

```
http://127.0.0.1:8000/predict/{feature}?features=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DqI0TQJI3vdU
```

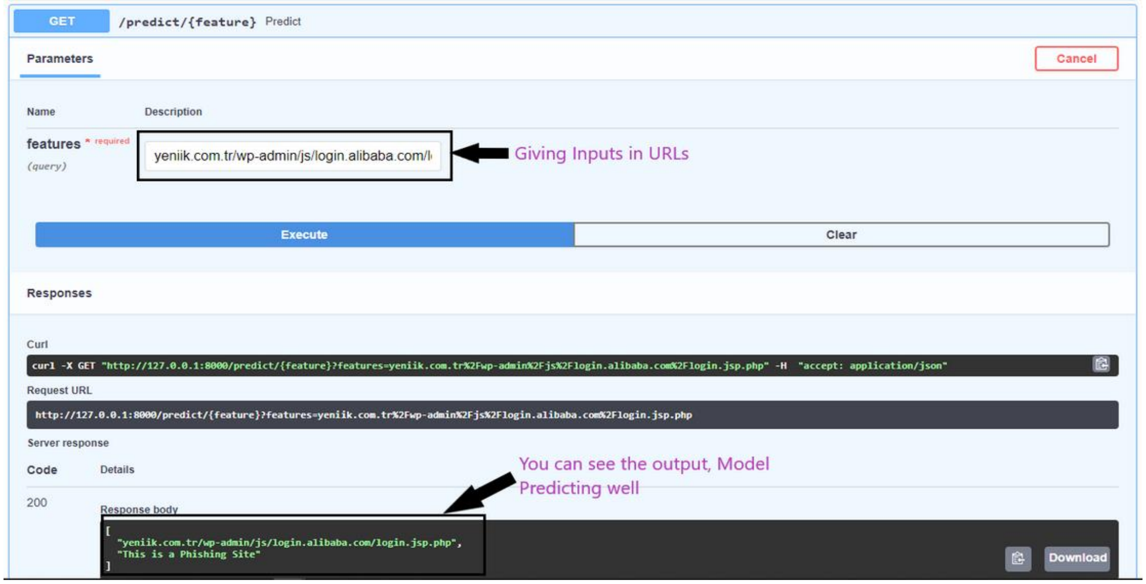
Server response

Code Details

200 Response body

```
[
  "https://www.youtube.com/watch?v=qI0TQJI3vdU",
  "This is not a Phishing Site"
]
```

You can see the Output, Model Predicting well.



features: yeniik.com.tr/wp-admin/js/login.alibaba.com/l

Giving Inputs in URLs

Curl

```
curl -X GET "http://127.0.0.1:8000/predict/{feature}?features=yeniik.com.tr%2Fwp-admin%2Fjs%2Flogin.alibaba.com%2Flogin.jsp.php" -H  "accept: application/json"
```

Request URL

```
http://127.0.0.1:8000/predict/{feature}?features=yeniik.com.tr%2Fwp-admin%2Fjs%2Flogin.alibaba.com%2Flogin.jsp.php
```

Server response

Code	Details
200	Response body

```
[
  "yeniik.com.tr/wp-admin/js/login.alibaba.com/login.jsp.php",
  "This is a Phishing Site"
]
```

You can see the output, Model Predicting well

6.1 Advantages and Disadvantages

A. Advantages

1. This system can be used by many E-commerce or other websites in order to have good customer relationship.
2. Use can make online payment securely.
3. With the help of this system user can also purchase products online without any hesitation.

B. Disadvantages

1. If internet connection fails, this system won't work.
2. An automation detection system reduces response time, identifies threats and can classify them with one click.

VIII. Conclusion

It is outstanding that a decent enemy of phishing apparatus ought to anticipate the phishing assaults in a decent timescale. We accept that the accessibility of a decent enemy of phishing device at a decent time scale is additionally imperative to build the extent of anticipating phishing sites. This apparatus ought to be improved continually through consistent retraining. As a matter of fact, the accessibility of crisp and cutting-edge preparing dataset which may gained utilizing our very own device will help us to retrain our model consistently and handle any adjustments in the highlights, which are influential in deciding the site class. Albeit neural system demonstrates its capacity to tackle a wide assortment of classification issues, the procedure of finding the ideal structure is very difficult, and much of the time, this structure is controlled by experimentation. Our model takes care of this issue via computerizing the way toward organizing a neural system conspire; hence, on the off chance that we construct an enemy of phishing model and for any reasons we have to refresh it, at that point our model will encourage this procedure, that is, since our model will mechanize the organizing procedure and will request scarcely any client defined parameters.

Acknowledgements

The completion of our project brings with it a sense of satisfaction, but it is never complete without those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine our completion of the project without guidance and neither can we succeed without acknowledging it. It is a great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the respected personalities.

REFERENCES

[1]. K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, ''A new hybrid ensemblefeature selection framework for machine learning-based phishing detection system,'' Inf. Sci.,vol. 484, pp. 153–166, May2019.

[2]. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, ''Machine learning based phishing detectionfrom URLs,'' Expert Syst. Appl., vol. 117, pp. 345–357, Mar. 2019.

[3]. R. M. Mohammad, F. Thabtah, and L. McCluskey, ''Tutorial and critical analysis of phishing Websites methods,''Comput. Sci. Rev., vol. 17, pp. 1–24, Aug. 2015.

[4]. K. L. Chiew, K. S. C. Yong, and C. L. Tan, ''A survey of phishing attacks: Their types, vectors and technical approaches,'' Expert Syst. Appl., vol. 106, pp. 1–20, Sep. 2018.

[5]. A. Aleroud and L. Zhou, ''Phishing environments, techniques, and countermeasures: A survey,''Comput. Secur., vol. 68, pp. 160–196, Jul. 2017.

[6]. V. Suganya, ''A review on phishing attacks and various anti phishing techniques,'' Int. J.Comput. Appl., vol. 139, no. 1, pp. 20–23, Apr. 2016.

[7]. N. Abdelhamid, A. Ayesh, and F. Thabtah, ''Phishing detection based associative classification data mining,'' Expert Syst. Appl., vol. 41, no. 13, pp. 5948–5959, Oct. 2014.

[8]. I. Qabajeh and F. Thabtah, ''An experimental study for assessing email classification attributes using feature selection methods,''in Proc. 3rd Int. Conf. Adv. Comput. Sci. Appl. Technol. (ACSAT), Dec. 2014, pp. 125–132.

[9]. R. S. Rao and A. R. Pais, ''Detection of phishing Websites using an efficient feature-based machine learning framework,'' Neural Comput. Appl., vol. 31, pp. 3851–3873, Jan. 2018.

[10]. W. Ali, ''Phishing Website detection based on supervised machine learning with wrapper features selection,'' Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 9, pp. 72–78, 2017.

[11]. APWG. Phishing Activity Trends Report 3rd Quarter 2019. Accessed: Mar. 21, 2020. [Online]. Available: https://docs.apwg.org/ reports/apwg_trends_report_q3_2019.pdf Dept. of

DOI: 10.48175/568