



# Phishing URLs Detection Using Machine Learning

Wend-Benedo Simeon Zongo<sup>(✉)</sup>, Boukary Kabore, and Ravirajsinh Sajubha Vaghela

Marwadi University, Rajkot, India

{zongowend-benedosimeon.114273,

kaboreboukarykaboretiga.114984}@marwadiuniversity.ac.in,

Ravirajsinh.vaghela@marwadieducation.edu.in

**Abstract.** Nowadays, internet user numbers are growing steadily, covering online services, and goods transactions. This growth can lead to the theft of users' private information for malicious purposes. Phishing is one technique that can cause users to be redirected to sites with malicious content and steal all of their information. The main purpose of phishing is to steal user identities such as online credentials, bank transaction details, etc. As technology advances, the mechanism of phishing attacks begins to take place, so to prevent it from happening, some mechanism anti-phishing is used to detect phishing links or URLs. Machine learning is the most solutions tools against phishing offensive, and with its algorithms, we can rank all content and determine whether it is phishing or not. We tested cross-validation as well as the correlation between features. Using Logistic Regression, we determined the importance of the features. Finally, we tested the Multinomial Naïve Baye classifier. We found that the Logistic Regression classifier had better accuracy for the best accuracy.

**Keywords:** Phishing · Domain name · Machine learning · URL · Classification models

## 1 Introduction

Internet services have brought immense changes to people's lives styles. Most online services are designed to connect users to membership systems and individual users must register and log in to receive these personalized services. For this reason, people must provide their personal information when entertaining this convenient and efficient service in a secure network. The environment, transmission, and storage of information are protected by network security technology. In addition, many cybercriminals use different methods to attack and steal personal information such as the case of phishing attacks.

Phishing is a technique used by most criminals via social digging of information and technical loopholes to steal consumers' secret information [1]. It is also a well-known, computer-based social engineering technique. Attackers are using disguised email addresses as a weapon to target large corporations to steal sensitive data. According

to some reports, as CISCO, in 2021, approximately 90% of data was breached due to phishing. Spear phishing is the most widely used type of phishing attack, comprising 65% of all phishing attacks. Studies carried out by Tessian in 2021 reveal that employees receive an average of 14 malicious emails per year. Cybercriminals use email scams because that way is simple, functional, and free. So, they encrypt all your email address information and send you emails in the name of a legitimate or original source.

To reduce this scourge which is a real threat to companies and individuals, approaches such as the anti-phishing extension for chrome and automatic detection of phishing links based on machine learning have been proposed. Anti-phishing chrome extension analyses all visited links to identify fake or right links related to their content [2]. Machine learning uses some algorithms to automatically analyzes and detects phishing URL with malicious content [3, 5, 6].

Machine learning is the ability of a computer to learn without being explicitly programmed [13]. Machine learning algorithms allow a system to automatically and repetitively learn from big data to predict or classify outcomes. The accuracy of predictions is determined through the quality and quantity of data. The learning process allows the machine to adjust over time to better adapt to the data, which improves performance. Consequently, an effective and efficient phishing detection approach is important to tackle the problem of phishing attacks [4]. This paper outlines different classification models of machine learning for phishing link detection such as logistic regression, decision trees, and natural language processing. Our work will be divided into 3 main parts to better analyze our document. As follow, Sect. 1: determine something related to the work. Section 2: Evoque our research methodology. Section 3: determine the results funds and analyze the best algorithms used.

## 2 Related Work

### 2.1 Literature Review

Phishing attack nowadays is increasing day by day. Since 2020 APWG was observing between 68 000 and 94 000 phishing attacks per month. But this number has tripled, APWG reported 316 747 unique phishing Web sites attacks in December 2021 which was the highest monthly total in APWG's reporting history during the period [8]. APWG recorded 1,025,968 total phishing attacks in Q1 2022. APWG counted 384,291 attacks in March 2022.

Regarding this report, in recent years, many documents and articles have been published demonstrating some methodologies and strategies to detect phishing domains or URLs. Many of them use a machine learning algorithm to detect malicious URLs. Classification model techniques are the better learning capabilities from cyber data [9] (Fig. 1).

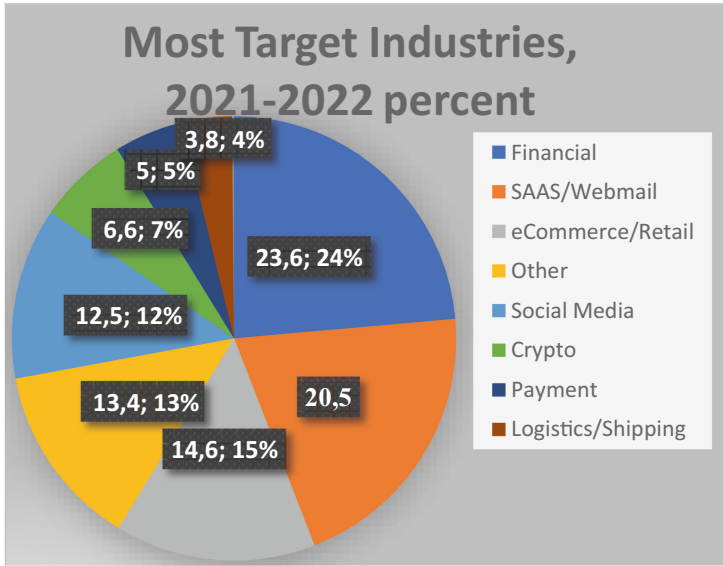


Fig. 1. APWG report in 2022

It mainly presents a machine learning-based approach to detect phishing websites in real-time, considering hybrid features based on URLs and hyperlinks to achieve high accuracy without relying on third-party systems.

Yadav, N., and Panda. [10] presented a mixed-selection model that combined both contents- and behavior-based methods to help identify the attacker using email headers. Manish Jain, Kanishk Rattan, Divya Sharma, Kriti Goel, and Nidhi Gupta have proposed a framework for detecting phishing sites using machine learning algorithms such as Naive Bayes Classifier, Random Forest, and Support Vector Machine. Among all these algorithms, Random Forest gives the most accuracy and this framework uses address bases, domains and HTML JS features to detect the legitimacy of the website.[11]

Suleiman Y. Yerima and Mohammed K. Alzaylae proposed a framework for detecting phishing websites using a deep learning approach [11]. They used the CNN (Convolutional Neural Network) model to achieve high accuracy. They used only the URL-based feature to detect the phishing site, it has 30 URL attributes. This approach has a better score than any other approach.

Weiwei Zhuang, Qingshan Jiang, and Tengke Xiong proposed an intelligent anti-phishing strategy model for phishing site detection [11]. It uses a heuristic URL detection module. It has a categorization module. It categorizes phishing as a bank, lottery, etc. It

uses a hierarchical clustering algorithm for phishing categorization. Rishikesh Mahajan and Irfan Siddavatam developed a phishing site detection system using machine learning algorithms such as decision trees, random forest, and support vector machine, where random forest gave the best accuracy [11].

## 2.2 URLs Descriptions

### URL Description

A URL (Uniform Resource Locator) is a unique address in a computer network, which allows to index of a data source. This data source or address can be an HTML page, an image, a document, etc... Each URL has a set structure [12].

`http://www.yourbank.in:80/Upload?key=values&keys=value#otherContent`

Structure and description of URLs

**Scheme:** HTTP or HTTPS. It represents the protocol

**Authority:** `www.yourbank.in:80`. It represents part of the domain name and port number which the protocol is able to use

**Resource Path: /Upload/.** It mentions files directories

**Parameters- Parameters:** `Key=values&keys=value`. Pieces of information in a query string of a URL

### Phishing URL

Attackers, usually change the subdomain name and path of the URL.

Example: `http://yourbank.in.account.yourbanks.it/` users

Structure:

Protocol: Http

Domain Name: `yourbanks.it`

Subdomain item 1: `yourbanks.in`

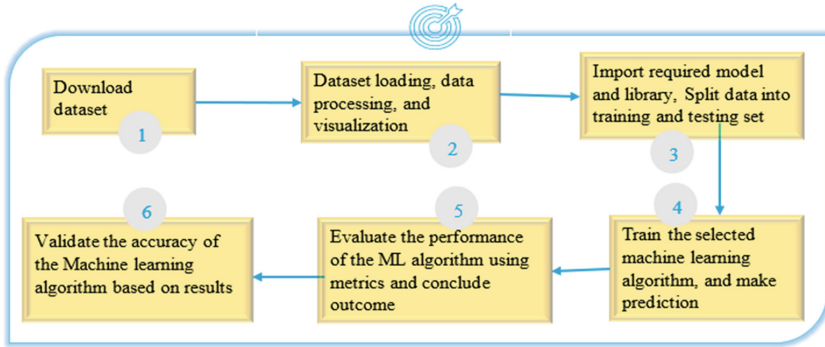
Subdomain item2: `account`

Path: `users`

Additionally, Attackers use Cybersquatting and Typo squatting techniques to tempt users. Example: `facebook.com`, they change one or many letters from the main, meaning the phishing link can be `facebool.com`.

## 3 Research Methodology

This section revolves around the different processes and methodologies used to achieve the result (Fig. 2).



**Fig. 2.** Research methodology

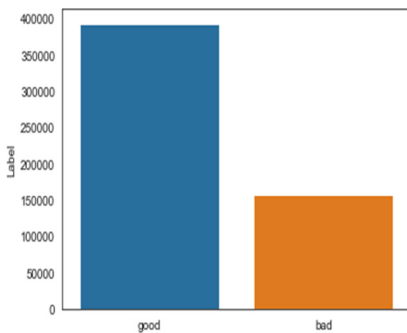
### Step 1: Dataset download

Dataset has been downloaded from Kaggle (Kaggle.com), a website containing many machine-learning datasets. The dataset is named “phishing\_site\_urls.csv”, and has 549346 entries and 2 columns. The prediction column is the Label which has 2 categories.

Bad: URLs contain malicious elements and these sites are phishing websites  
 Good: URLs don’t contain malicious elements these sites are good websites

### Step 2: Dataset loading and processing

Panda’s library has been imported to load the dataset. The method *read\_csv()* is used to create the Data Frame from the CSV file (dataset). Methods such as *IsNull()*, and *sum()* have been used for dataset processing. Indeed *IsNull().sum()* counts all null values of each column. We used the *value\_count()* method on the label column to find the number of good and bad URLs. As a result, we found 392924 for good and 156422 for bad. Seaborn from matplotlib is used to visualize the data of the target column (Label).



**Fig. 3.** Dataset visualisation based on label categories

Count Vectorizer and tokenizer are used to prepare the predictive columns. We imported Regex Tokenizer and Snowball Stemmer functions from the *nlk* library. *Regex Tokenizer* through regular expressions splits a string into substrings and allows special characters to be removed from the URL using the Tokenize method. As for *Snowball Stemmer*, it reduces the words of the URL to their base root. We have created columns for tokenizing named “text\_tokenized” text and snowball text named “text\_stemmed”. After we have combined both into a single column named “text\_send” (Fig. 3).

A function has been created to allow easy data visualization. This function uses as a library matplotlib, word cloud, STOPWORD, and Image color generator. Matplotlib provides object- oriented API for embedding plot info into applications. Words Cloud is a data visualization technique used for representing text data. It can be represented in the following picture.

Important tools, such as Selenium web driver, have been used to visualize internal links. It offers features for browsing web pages, user input, etc. To add, it scraps dynamic websites for testing. For use (Fig. 4):



**Fig. 4.** Word Cloud data visualisation

Download chromedriver.exe corresponding to the same version of your navigation.

Set up the Chrome driver by creating a list of URLs.

List all links to the created list Create an empty list that will append all links containing each website.

Use the BeautifulSoup library to extract only hyperlinks that are relevant to Google: links only with '<a>' tags with href attributes.

### **Step 3: Model library, splitting data into training and testing sets**

Some libraries are imported for data predictions:

Count vectorizer is a python library that allows the conversion of a collection of text documents to a matrix of token counts. It comes from sci-kit-learn. The *fit transform* method is called to transform all text that we tokenized and stemmed.

*Train test split* allows splitting the dataset into train and test data. This function is imported from the sci-kit-learn library.

*Logistic Regression* is used to predict the likelihood of a categorical dependent variable. In other words, the logistic regression model predicts  $P(Y=1)$  given  $X$ . We appealed the Naive Bayes multinomial classifier to predict the tags of the text of our label with the greatest chance. It is well known for discrete feature classification such as spam filtering, sentiment analysis, and text Classification. *Classification report and accuracy score* are used to give all reports about metric as recall, f1 score, prediction, etc...

*Confusion metric* is used to give all info of actual prediction.

#### **Step 4: Train the selected machine learning algorithm, and make a prediction**

Function `train_test_split()` is used for data training and testing. The column Label is used as a Y value for prediction. For the x value, we used the fit transform method from the count vectorizer library to transform all text that has been tokenized of `text_sent` columns. The x value is named “*feature*”.

## **4 Result and Analysis**

### **4.1 Result**

The results prediction of our two algorithms (Logistic Regression and Multinomial Naïve Baye) are returned in the below table (Table 1).

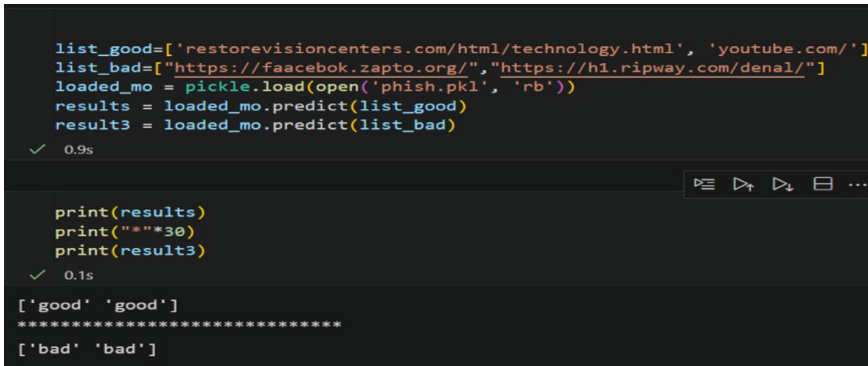
**Table 1.** Evaluation of algorithms score

Algorithm	Function	Result percent
Logistic regression	Testing accuracy (score)	96%
	Training accuracy	97%
Multinomial Naïve Bayes	Testing accuracy (score)	95%
	Training accuracy	96%

Based on preview results we see that logistic regression gives a better prediction of 96%.

### **4.2 Analysis**

The pipeline is used with Logistic Regression to analyze the model of classification. We use the *make pipeline* function to combine all processor techniques for predicting URLs real. We graph file a database named “phishing. Pkl” for testing with other links. The output predicts the stat of URLs (Fig. 5).



```

list_good=['restorevisioncenters.com/html/technology.html', 'youtube.com/']
list_bad=["https://faacebok.zapto.org/","https://hl.ripway.com/denal/"]
loaded_mo = pickle.load(open('phish.pkl', 'rb'))
results = loaded_mo.predict(list_good)
result3 = loaded_mo.predict(list_bad)

✓ 0.9s

print(results)
print("*"*30)
print(result3)

✓ 0.1s

['good' 'good']
*****
['bad' 'bad']

```

Fig. 5. Data prediction using graph file.

## 5 Conclusion

This paper presents a mechanism used to detect phishing URLs. We have managed to use machine learning as a more powerful tool to solve this problem. Two machine learning algorithms were used to predict the data. Among them, Logistic Regression gives a better score of about 96%. This classification model is used to predict URLs outside the dataset. The results predict the status of the website. As a perspective, we will link the algorithms to a browser for visibility of the prediction results.

## References

1. Hong, J., Kim, T., Liu, J., Park, N., Kim, S.-W.: Phishing URL detection with lexical features and blacklisted domains. In: Jajodia, S., Cybenko, G., Subrahmanian, V.S., Swarup, V., Wang, C., Wellman, M. (eds.) Adaptive autonomous secure cyber systems, pp. 253–267. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-33432-1\\_12](https://doi.org/10.1007/978-3-030-33432-1_12)
2. Sharma, H., Meenakshi, E., Bhatia, S.K.: A comparative analysis and awareness survey of phishing detection tools. In: Proceedings of the 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT), pp. 1437–1442. IEEE (2017)
3. Dutta, A.K.: Detecting phishing websites using machine learning techniques. PLoS ONE **16**(10), e0258361 (2021)
4. Assegie, T.A.: K-nearest neighbor based URL identification model for phishing attack detection. Indian J. Artif. Intell. Neural Networking (IJAINN) (2021)
5. Homayoun, S., Hageman, K., Afzal-Houshmand, S., Jensen, C.D., Pedersen, J.M.: Detecting ambiguous phishing certificates using machine learning. In: Proceedings of the 2022 International Conference on Information Networking (ICOIN), pp. 1–6. IEEE (2022)
6. Rather, D., Mann, S.: Detection of E-mail phishing attacks—using machine learning and deep learning. Int. J. Comput. Appl. **183**, 1–7 (2022)
7. Butt, U.A., Amin, R., Aldabbas, H., Mohan, S., Alouffi, B., Ahmadian, A.: Cloud-based email phishing attack using machine and deep learning algorithms. Complex Intell. Syst., 1–28 (2022)
8. APWG: Phishing activity trends report, 1st quarter 2022, June 2022



9. Das Gupta, S., Shahriar, K.T., Alqahtani, H., Alsalman, D., Sarker, I.H.: Modeling hybrid feature-based phishing websites detection using machine learning techniques. *Ann. Data Sci.* **1874**, 1–26 (2022). <https://doi.org/10.1007/s40745-022-00379-8>
10. Yadav, N., Panda, S.P.: Feature selection for email phishing detection using machine learning. In: Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds.) *International Conference on Innovative Computing and Communications. AISC*, vol. 1388, pp. 365–378. Springer, Singapore (2022). [https://doi.org/10.1007/978-981-16-2597-8\\_31](https://doi.org/10.1007/978-981-16-2597-8_31)
11. Kureel, V.K., Maurya, S., Shaikh, A., Tiwari, S., Nagmote, S.: Phishing website detection using machine learning, 7421. Journal homepage: [www.ijrpr.com](http://www.ijrpr.com). ISSN: 2582
12. Chaudhari, M.S.S., Gujar, S.N. and Jummani, F., Detection of phishing web as an attack: a comprehensive analysis of machine learning algorithms on phishing dataset (2022)
13. Ott, M.A.: Bias in, bias out: ethical considerations for the application of machine learning in pediatrics. *J. Pediatrics* (2022). <https://www.kaggle.com/datasets/taruntiwarihp/phishing-site-urls>, GitHub link: <https://github.com/phishing-ml/phishing-ml>