Phishing Classifier Using Machine Learning

[1] Mohd Saifulla, [2] Syed Mahmood, [3] Mohd Subhan Khan,

[4] Mr. Mohammed Rahmat Ali

[1] BE Student, Dept. of Computer Science Engineering, ISL Engineering College

[2] BE Student, Dept. of Computer Science Engineering, ISL Engineering College

[3] BE Student, Dept. of Computer Science Engineering, ISL Engineering College

[4] Assistant Professor, Dept. of Computer Science Engineering, ISL Engineering College

Article Info
Page Number: 1292-1299
Publication Issue:
Vol. 72 No. 1 (2023)

Article History
Article Received: 15 October 2022
Revised: 24 November 2022
Accepted: 18 December 2022

ABSTRACT

This paper proposes a website phishing classifier using machine learning techniques. The proposed classifier uses a feature-based approach to extract relevant features from website URLs and contents. The extracted features are then used to train a classification model based on various machine learning algorithms such as Random Forest, Support Vector Machine, and Naive Bayes. The performance of the proposed classifier is evaluated using a publicly available dataset of phishing and legitimate websites. The experimental results show that the proposed classifier achieves high accuracy, precision, and recall in detecting phishing websites. The proposed classifier can be used as an effective tool to detect phishing websites in real-time and prevent users from falling victim to phishing attacks.

Keywords – Website phishing, Machine Learning, Classifier.

1. Introduction

A classifier to predict whether a website is phishing, based on a specified set of predictors. Phishing sites are a type of cybercrime attack aimed at stealing sensitive online user information, including login Access to your data, and bank details. Attackers trick users by presenting fake websites that appear to be legitimate or trustworthy for obtaining valuable data.

Various solutions have been proposed to detect and prevent phishing attacks. Sites such as heuristics, blacklists, whitelists, and machine learning (ML) Techniques. This study examines the latest ML techniques for detection Identify phishing websites and identify possible solutions to the problem. The majority of existing ML techniques are based on traditional techniques such as random forests(RF), Support Vector Machines (SVM), Naïve Bayes (NB), and Ada Boosting [1][6]. However, deep learning-based techniques have shown better performance. A traditional ML method for detecting phishing websites. some challenges when faced with ML methods include overfitting, low accuracy, and poor training data. This research suggests that Internet users should be aware of phishing attacks. Avoid falling victim to them and suggest developing automated solutions to detect phishing websites[2].

One of the challenges our research faced was the lack of availability of a reliable training record. Virtually all researchers in this field face this challenge [2]. Although many articles on predicting phishing websites use data mining techniques, authoritative training datasets may not be published as there is no consensus in the literature about what characterizes features.

While addressing this issue categorizing the website features by which the legitimacy of a website can be decided and also helps in training the model easily[3].

- [4] Characteristics or attributes of a website that are based on the address bar or URL of the website.
- Attributes of a website that are considered abnormal or unusual can be used to detect potentially malicious websites or phishing attempts.
- Based on the HTML and JavaScript code used to create the website, which can be used to identify specific elements or functionality of the website.
- An attribute of a website based on the domain name of the website, and can be used to distinguish between legitimate and malicious websites.

1.1 Scope of the project:

The primary goal of the project is to identify phishing websites through various approaches and make the internet safe, this is done by identifying the phishing websites across the internet. This further helps in preventing the loss of confidential data of a user across a website and also preventing scams etc.

2 Literature Survey

There are a lot of factors that affect the detection of the phishing website, based on the characteristics of the URL and it also includes psychological and social factors let us understand this by taking an example of the IP address

2.1 IP Address

If the URL uses an IP address instead of a domain name for instance ."http://125.98.3.123/fake.html" users can rest assured that someone is trying to steal their personal information. The IP address may also be converted to his hexadecimal code, as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".rule:

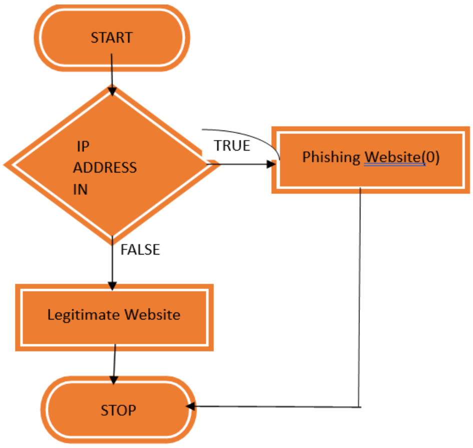


Fig 1: Flow chart

The features of a website demonstrate its legitimacy, it can be used to Identify phishing websites based on their features[4][8]. These factors help in training the model and easily classifying the data based on the features which help us to train the model for phishing detection of the website. Below are how the features are classified.

Here are some of the features of the website which are used to train the data by using the approach of [0, 1, -1] This helps to train the model easily, below is an example of how the feature is used to classify the website.

2.1.1. Long URL to hide Suspicious part

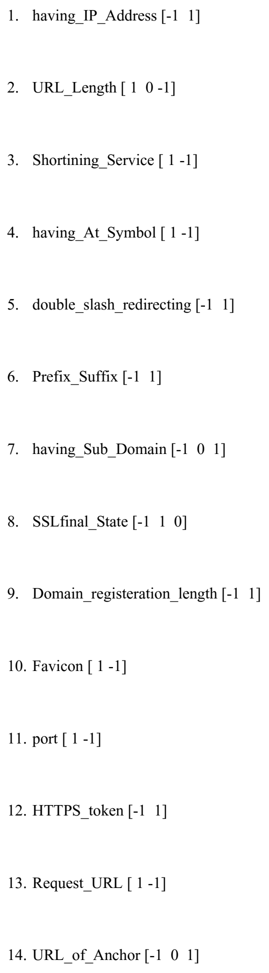
Rule: IF{

$URL\ length < 56 \rightarrow feature =$ Legitimate

$else\ if\ URL\ length \geq 51\ and \leq 85 \rightarrow feature = Suspicious$

$otherwise \rightarrow feature =$ Phishing

Below are some examples how features are used



1. having_IP_Address [-1 1]
2. URL_Length [1 0 -1]
3. Shortining_Service [1 -1]
4. having_At_Symbol [1 -1]
5. double_slash_redirecting [-1 1]
6. Prefix_Suffix [-1 1]
7. having_Sub_Domain [-1 0 1]
8. SSLfinal_State [-1 1 0]
9. Domain_registeration_length [-1 1]
10. Favicon [1 -1]
11. port [1 -1]
12. HTTPS_token [-1 1]
13. Request_URL [1 -1]
14. URL_of_Anchor [-1 0 1]

Fig 2: Features of websites

This paper uses various approaches of machine learning algorithms used for phishing detection, including decision trees, support vector machines, k-nearest neighbors, and neural networks.

In a study, various machine algorithms were used and trained to detect phishing websites but due to a lack of data availability, the result generated was not accurate using multiple features and training the model the system proposed by us can detect the accuracy way higher when compared to the previous study.

When the algorithms are used to detect phishing websites multiple features and algorithms are used to predict the outcome so that it should be accurate.[8]

Data Validation :

Name Validation- We validate the name of the files based on the given name in the schema file. We have created a regex pattern per the name given in the schema file for validation. After validating the pattern in the name, we check for the length of the date in the file name as well as the length of time in the file name. If all the values are as per requirement, we move such files to "Good_Data_Folder" else we move such files to "Bad_Data_Folder."

2. Number of Columns - We validate the number of columns present in the files, and if it doesn't match the value given in the schema file, then the file is moved to "Bad_Data_Folder."

3. Name of Columns - The columns' names are validated and should be the same as given in the schema file. If not, then the file is moved to "Bad_Data_Folder".

4. The datatype of columns - The datatype of columns is given in the schema file. This is validated when we insert the files into Database. If the datatype is wrong, then the file is moved to "Bad_Data_Folder".

5. Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to "Bad_Data_Folder"

3 Methodology

3.1 Model Training

1) Data Export from Db - The data in a stored database is exported as a CSV file to be used for model training.

2) Data Preprocessing

a) Replace the invalid values with numpy “nan” so we can use imputer on such values.

b) Check for null values in the columns. If present, impute the null values using the KNN imputer.

3) Clustering - KMeans algorithm is used to create clusters in the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using the "KneeLocator" function.

The idea behind clustering is to implement different algorithms. To train data in different clusters. The K-means model is trained over preprocessed data and the model is saved for further use in prediction.

4) Model Selection - After clusters are created, we find the best model for each cluster. We are using two algorithms, "SVM" and "XGBoost". For each cluster, both algorithms are passed with the best parameters derived from GridSearch.

To train the model a set amount of data was used to train it which can be improved future so that the algorithm produces better results.

3.2 Prediction

While selecting a legitimate website all the predictions are taken into consideration and the outcome of the website is decided whether the website is phishing or legitimate.

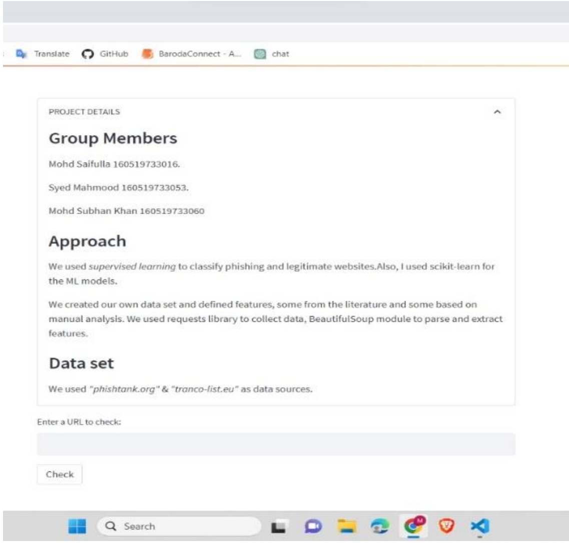
It was observed that while training the data the more accurate the data is provided for the training of the model the accuracy of the answers is increased and the model is improved, features of a website are categorized and presented to the training model.

The database used in this model is designed in such a way that the features are categorized based on 1,0,-1 which declares the legitimacy where 1 declares as yes 0 is neutral, and -1 as false.

	ML Model	Accuracy	f1_score	Recall	Precision
1	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
2	CatBoost Classifier	0.972	0.975	0.994	0.989
3	XGBoost Classifier	0.969	0.973	0.993	0.984
4	Multi-layer Perceptron	0.969	0.973	0.995	0.981
5	Random Forest	0.967	0.971	0.993	0.990
6	Support Vector Machine	0.964	0.968	0.980	0.965
7	Decision Tree	0.960	0.964	0.991	0.993
8	K-Nearest Neighbors	0.956	0.961	0.991	0.989
9	Logistic Regression	0.934	0.941	0.943	0.927
10	Naive Bayes Classifier	0.605	0.454	0.292	0.997

Table 1: AUC Score of the algorithms

4 Implementation



Group Members

Mohd Saifulla 160519733016.

Syed Mahmood 160519733053.

Mohd Subhan Khan 160519733060

Approach

We used *supervised learning* to classify phishing and legitimate websites.Also, I used scikit-learn for the ML models.

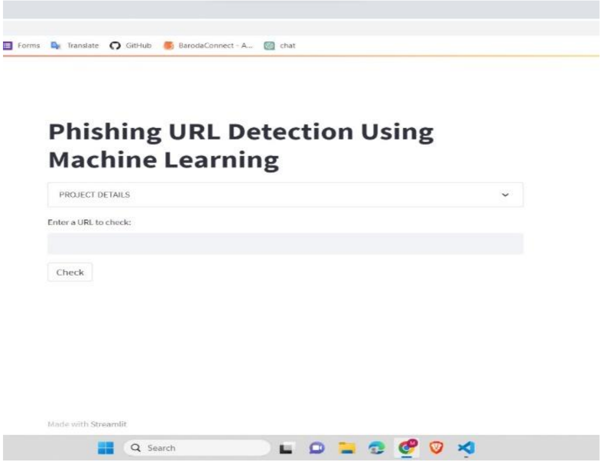
We created our own data set and defined features, some from the literature and some based on manual analysis. We used requests library to collect data, BeautifulSoup module to parse and extract features.

Data set

We used "*phishtank.org*" & "*tranco-list.eu*" as data sources.

Enter a URL to check:

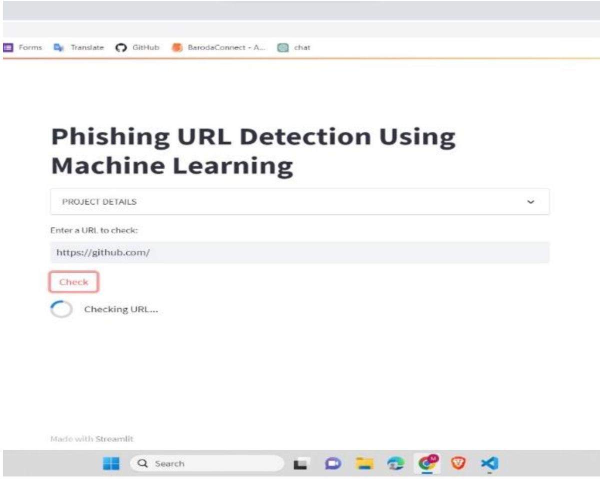
Fig 3: User Interface



Phishing URL Detection Using Machine Learning

Enter a URL to check:

Fig 4: Main Page



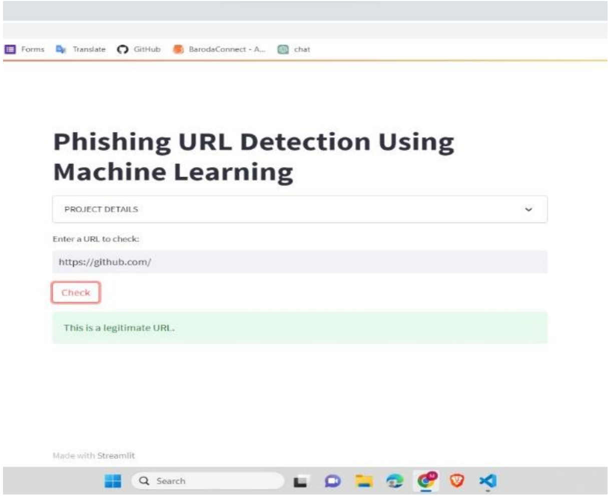
Phishing URL Detection Using Machine Learning

Enter a URL to check:

https://github.com/

Checking URL...

Fig 5: URL Detection



Phishing URL Detection Using Machine Learning

Enter a URL to check:

https://github.com/

This is a legitimate URL.

Made with Streamlit

Fig 6: Phishing or Legimate

4 Conclusion and Future Enhancement

In conclusion, as businesses are being digitalized, a lot of people are trying to bring their businesses online to buy and sell products, as important credentials are saved on the servers or on a user's account at the company's end if an attacker tries to send a phishing link to the user and steals his card or other personal details and sells them online or misuses them. To prevent such things, various machine learning algorithms and models can be trained and introduced to prevent such attacks, so that before visiting a website, users should check its legitimacy and secure its data online.

While doing this research and experimenting with the data, we observed that it can be further enhanced by using valid or more accurate databases or data, as we used the data generated from our end at a low scale and also used some extensions and APIs to process the data available online and train the model more precisely, such that the output generated should be accurate and reliable and can prevent online phishing attacks. This can help prevent the loss of data, knowingly or unknowingly, on the internet, as a lot of people find it difficult to differentiate between a legitimate and phishing website.

5 References

1. Siddharth, K. M., & Salankar, S. S. (2020). An intelligent anti-phishing model using enhanced ML techniques. Journal of Ambient Intelligence and Humanized Computing, 11, 2997-3014. doi: 10.1007/s12652-019-01475-w
2. Dhabliya, D. (2019). Security analysis of password schemes using virtual environment. International Journal of Advanced Science and Technology, 28(20), 1334-1339. Retrieved from www.scopus.com
3. Dhabliya, D., & Dhabliya, R. (2019). Key characteristics and components of cloud computing. International Journal of Control and Automation, 12(6 Special Issue), 12-18. Retrieved from www.scopus.com

4. Dhabliya, D., & Parvez, A. (2019). Protocol and its benefits for secure shell. International Journal of Control and Automation, 12(6 Special Issue), 19-23. Retrieved from www.scopus.com
5. Dhabliya, D., & Sharma, R. (2019). Cloud computing based mobile devices for distributed computing. International Journal of Control and Automation, 12(6 Special Issue), 1-4. doi:10.33832/ijca.2019.12.6.01
6. Dhabliya, D., Soundararajan, R., Selvarasu, P., Balasubramaniam, M. S., Rajawat, A. S., Goyal, S. B., . . . Suciu, G. (2022). Energy-efficient network protocols and resilient data transmission schemes for wireless sensor Networks—An experimental survey. Energies, 15(23) doi:10.3390/en15238883
7. Menon, D. S., & Pai, N. P. (2017). Phishing website detection using machine learning algorithms. International Journal of Computer Applications, 175(3), 10-14. doi: 10.5120/ijca2017914182
8. Balogun, A. A., Olabiyisi, O. D., & Ajayi, O. S. (2020). Phishing websites detection using machine learning. Journal of Intelligent & Fuzzy Systems, 39(1), 879-890. doi: 10.3233/JIFS-179580
9. Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey (2020) Phishing website features published at the University of Huddersfield Huddersfield, UK.
10. Asadullah Safi (SLR). (2023, January 11). *A systematic literature review on phishing website detection techniques*. Journal of King Saud University - Computer and Information Sciences.
11. “PhishStorm: Detecting Phishing With Streaming Analytics,” IEEE Transactions on Network and Service Management, vol. 11 , issue: 4 .
12. Huang, Yongjie; Yang, Qiping; Qin, Jinghui; Wen, Wushao (2019). [IEEE 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering.
13. Zongo, WB.S., Kabore, B., Vaghela, R.S. (2022). Phishing URLs Detection Using Machine Learning. In: Rajagopal, S., Faruki, P., Popat, K. (eds) Advancements in Smart Computing and Information Security. ASCIS 2022. Communications in Computer and Information Science, vol 1760. Springer, Cham.
14. 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)
15. Muskaan Tabasum, Fouzia Sultana, Sumayya Maheen Unnisa and Dr. Pathan Ahmed Khan. (n.d.). CYBERBULLYING DETECTION USING NAIVE BAYES ALGORITHM. Google Docs. Retrieved May10,2023,fromhttps://drive.google.com/file/d/18kCQf5ItxGPz0eqrYsJvxEulF06lUlN/view?usp=drivesdk
16. Ali, R., Ishaqui, S. Y. A., Shareef, M. F., & Khan, P. A. (n.d.). Machine learning inspired code word selection for dual connectivity in 5G user-centric ultra- dense networks. Ijrar.org. Retrieved May 10, 2023, from https://www.ijrar.org/papers/IJRAR22B2442.pdf