INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 3 - V8I3-1404)
Available online at: https://www.ijariit.com

Detection of phishing websites using a Machine Learning approach

Koushik Kumar Reddy
koushikkumar1807@gmail.com
Jain University, Kanakapura, Bengaluru

E. Surendra
18btrce007@jainuniversity.ac.in
Jain University, Kanakapura, Bengaluru

M. Sai Nikhil Gowd
18btrce020@jainuniversity.ac.in
Jain University, Kanakapura, Bengaluru

N. Jaya Pradeep Reddy
18btrce021@jainuniversity.ac.in
Jain University, Kanakapura, Bengaluru

Narasimhayya B. E.
e.narasimhayya@jainuniversity.ac.in
Jain University, Kanakapura, Bengaluru

ABSTRACT

Advances in Internet and cloud technology have responded in a major expansion in electronic trade, in which consumers conduct online purchases and deals, in recent times. This expansion results in unlawful access to sensitive information held by druggies, as well as damage to a company's coffers. Phishing is a well- known attack that deceives junkies into viewing vicious content in order to steal their particular information. utmost phishing webpages act identical to licit webpages in terms of website interface and universal resource position (URL). Colorful styles for relating phishing websites have been proposed, including blacklisting, heuristics, and so on. still, the number of victims is adding exponentially as a result of ineffective security technologies. Phishing assaults are more likely on the Internet because of its anonymous and limited nature.

Keywords-Mutual Info, Lexical Features, Logistic Regression, Random Forest, Naive Bayes, F1 Score and Accuracy.

I. INTRODUCTION

Phishing is a deceptive tactic that involves the use of social and technological wile to steal a client's identity and fiscal information. We do the maturity of our work on digital platforms in our diurnal lives. In numerous ways, having a computer and access to the internet makes our work and particular lives easier. It enables us to perform deals and operations in fields including trade, health, education, communication, banking, aeronautics, exploration, engineering, entertainment, and public services in a timely manner. With the advancement of mobile and wireless technologies, users who require access to a local network can now effortlessly connect to the Internet from anywhere and at any time. Although this arrangement is quite convenient, it has highlighted major information security flaws. As a result, the necessity for cyberspace users to take precautions against potential cyber-attacks has arisen. Fraud, forgery, coercion, shakedown, hacking, service blocking, virus software, illicit digital materials, and social engineering are all common targets of these attacks. The average cost of an assault in 2019 (depending on the scale of the attack) is between $ 108K and $ 1.4 billion, according to Kaspersky's research. Furthermore, the total cost of worldwide security goods and services is estimated to be approximately $ 124 billion. Phishing assaults are the most common and dangerous of these types of attacks. It results in monetary and intangible losses. In phishing assaults, the technique of contacting target individuals has changed.

In 2019, the average fiscal cost of a data breach as a result of phishing attempts was$3.86 million, while the projected cost of BEC (Business Dispatch concession) terms was$ 12 billion. likewise, it's estimated that 15 of those who are assaulted have at least one other target. As a result, it's reasonable to prognosticate that phishing assaults will continue in the coming times. So, using machine literacy ways and algorithms similar as Logistic Retrogression, KNN, SVC, Random Forest, Decision Tree, XGB Classifier, and Nave Bayes, we proposed a system to prognosticate Phishing Websites grounded on colorful parameters uprooted from the website link entered by the stoner in the frontal end.

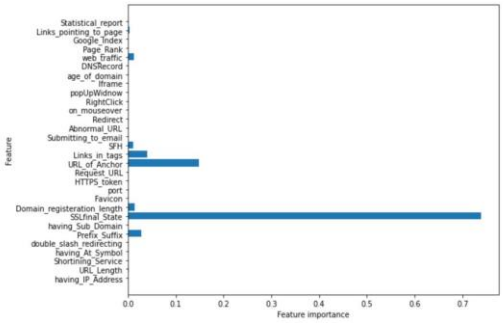

Figure 1: Lexical Features Considered

II. LITERATURE REVIEW

Machine Learning Technique for Phishing Detection: T. Mahmood, M. W. Nisar, and T. Nazir: J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir:
This study presents a phishing detection system technique for detecting blacklisted URLs, also known as phishing websites, so that users can be informed while browsing or visiting a certain website. As a result, it may be used for identification and verification, and it can also be used to protect people from being duped. In compared to other applications, the system has a lot of functions. It has unique capabilities such as recording blacklisted URLs straight from the browser to check the website's legitimacy, informing users on banned websites as they try to access through popup, and also notifying via email. This technology will aid users in being aware of their surroundings.

Machine learning for Phishing Website Discovery M.B.H. Frej, D. Sabyrov, A. Shaikhyn,F. Amsaad, and A. OunA. Razaque, M.B.H. Frej, D. Sabyrov, A. Shaikhyn, F. Amsaad, and A. Oun By designing an extension for the Google Chrome web cyber surfer, we contribute to the result of the phishing problem in this exploration. We utilised JavaScript PL in the construction of this functionality. A blend of Blacklisting and semantic analysis tools were employed to descry and baffle the fishing assault. In addition, a phishing point database is created, and the textbook, links, prints, and other data on the point are estimated for pattern recognition. Eventually, our recommended result was put to the test and compared to other options. The findings show that our proposed strategy is able of effectively dealing with the phishing problem. Limitations devoted to all aspects of the job.

A regular overview on Phishing Detection Along With an Organized way to Construct an Anti-Phishing Framework S. Patil and S. Dhage.
Phishing is a security attack to acquire particular information like watchwords, credit card details or other account details of a stoner by means of websites or emails. Phishing websites look analogous to the licit bones which make it delicate for a nonprofessional to separate between them. As per the reports of Anti Phishing Working Group(APWG) published in December 2018, phishing against banking services and payment processor was high. nearly all the fraudulent URLs use HTTPS and use redirects to avoid getting detected. This paper presents a focused literature check of styles available to descry phishing websites. A relative study of the in- use anti-phishing tools was fulfilled and their limitations were conceded. We analysed the URL-grounded features used in the history to ameliorate their delineations as per the current script which is our major donation. Also, Machine Learning Based Approach To descry Phishing Attacks a step wise procedure of designing an anti-phishing model is bandied to construct an effective frame which adds to our donation. compliances made out of this study are stated along with recommendations on being systems.
Limitations Anti phishing tools are hamstrung to descry all the phishing websites.

M. Korkmaz, O.K. Sahingoz, and B.Diri Discovery of Phishing Websites Using Machine Learning. They developed a machine literacy- grounded phishing discovery system in this report, which used eight different algorithms to assay URLs and three distinct datasets to compare the findings to former exploration. The experimental findings show that the suggested models operate exceptionally well, with a high rate of success. We used machine literacy styles in this study to produce a phishing discovery system. The suggested systems are put to the test using some current datasets from the literature, and the results are compared to the most recent workshop. The results of the comparison reveal that the suggested styles ameliorate phishing discovery effectiveness and achieve high delicacy rates, in the coming times.

III. METHODOLOGY

Detection of plant phishing URL, steps are carried out as shown in below figure 2:

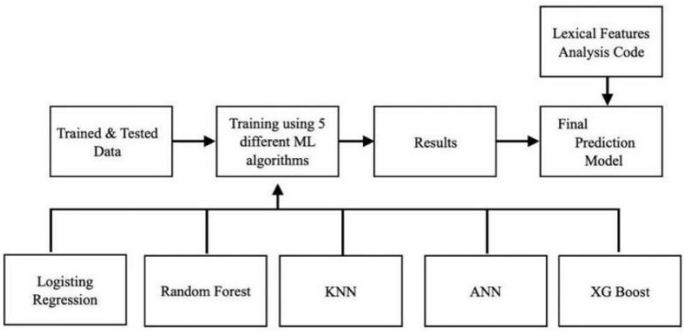

Figure 2: Proposed Methodology

To make it easier to develop and test the complicated project, we've divided it into two sections.

Phase 1:
- The first step is to collect a Phishing Website Dataset.
- Various machine learning algorithms are used to load and pre-process the dataset.
- Training and testing data are separated from the pre-processed data.
- Machine learning methods such as Proactive Phishing URL websites, Logistic Regression, KNN, SVC, Random Forest, Decision Tree, Naive Bayes, and XGB Classifier are used to create the prediction model.

Phase 2:
- The model is trained using the training dataset, and when it has been properly trained, it must be tested.
- The accuracy of the trained model is calculated after it has been evaluated using the testing dataset.
- Our final prediction model is based on the algorithm with the highest accuracy.

IV. RESULTS

The results are generated using Google colab software tool as shown in below figures:

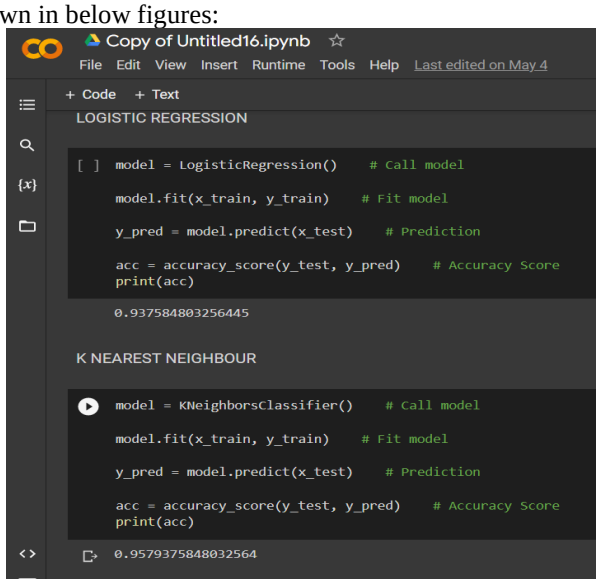

Figure 3(a): Training Dataset

Copy of Untitled16.ipynb

File Edit View Insert Runtime Tools Help Last edited on May 4

+ Code + Text

ARTIFICIAL NEURAL NETWORKS

```
[ ]  model = MLPClassifier()     # Call model

     model.fit(x_train, y_train)     # Fit model

     y_pred = model.predict(x_test)     # Prediction

     acc = accuracy_score(y_test, y_pred)
     print(acc)
```

```
     0.9642695612844867
```

RANDOM FOREST NETWORK

```
[ ]  model = RandomForestClassifier()     # Call model

     model.fit(x_train, y_train)     # Fit model

     y_pred = model.predict(x_test)     # Prediction

     acc = accuracy_score(y_test, y_pred)     # Accuracy Score
     print(acc)
```

```
     0.9764812302125735
```

Figure 3(b): Training Dataset and Accuracy

Performance Metrics:

Sl. no	Algorithm Used	Accuracy in %
1	Logistic Regression	93.4722
2	Random forest	97.5124
3	K-Nearest neighbor classifier	95.7033
4	Artificial Neural Network	97.2410
5	XG-Boost	96.4322

V. CONCLUSION

The proposed study's major thing is to stress the phishing strategy in the environment of bracket, where phishing websites are defined as websites that are automatically classified into a preset set of class values grounded on numerous attributes and the class variable. Website features are used by ML- grounded phishing tactics to acquire information that may be used to classify websites and descry phishing spots. Although phishing can-not be fully abolished, it may be dropped by strengthening targeted anti-phishing processes and tactics and educate the public on how to honor and identify bogus phishing websites.

ACKNOWLEDGMENT

The Honorable president, Chancellor ,Vice-Chancellor, Registrar, Director, and other staff members of the CSE department, School of Engineering & Technology, Jain University, are thankful for their strong provocation, support, and stimulant in all aspects of this paper's publication.

REFERENCES

[1] "Phishing Detection Using Machine Learning Technique," 2020 First International Conference on Smart Systems and Emerging Technologies (SMARTTECH), pp. 43-46, doi: 10.1109/SMART-TECH49988.2020.00026. J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference on Smart Systems and Emerging Technologies (SMARTTECH)

[2] "Detecting Phishing Websites Using Machine Learning," 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), 2020, pp. 111-114, doi: 10.1109/CSPA48992.2020.906872. M. H. Alkawaz, S. J. Steven, and A. I. Hajamydeen, "Detecting Phishing Websites Using Machine Learning," 16th IEEE International Colloquium on Signal Processing & It.

[3] "Detection and Prevention of Phishing Websites Using Machine Learning Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697412. V. Patil, P. Thakkar, C. Shah, T. Bhat, and S. P. Godse.

[4] "Phishing Website Detection Based on Machine Learning Algorithm," 2020 International Conference on Computing and Data Science (CDS), pp. 293-298, doi: 10.1109/CDS49703.2020.00064. W. Bai.

[5] "Detection of Phishing Websites Using Machine Learning," 2020 IEEE Cloud Summit, pp. 103-107, doi: 10.1109/IEEECloudSummit48914.2020.00022. A. Razaque, M. B. H. Frej, D. Sabyrov, A. Shaikhyn, F. Amsaad, and A. Oun.

[6] "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 384-389, doi:

10.1109/ICEECCOT46775.2019.9114695. M. M. Vilas, K. P. Ghansham, S. P. Jaypralash, and P. Shila.

[7] "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), pp. 1-6, doi: 10.1109/CAIS.2019.8769571. A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani,

