Acta Scientific COMPUTER SCIENCES

Volume 5 Issue 7 July 2023

Research Article

Phishing Detection for Covid-19 Theme-Based Email and Weblinks Using Machine Learning

Usman Ali* and Gul Bano

Department of Software Engineering, Mehran University of Engineering and Technology, Pakistan

***Corresponding Author:** Usman Ali, Department of Software Engineering, Mehran University of Engineering and Technology, Pakistan.

Received: May 09, 2023

Published: June 10, 2023

Abstract

During the COVID-19 pandemic, phishing frauds became more prevalent as the victim was easily deceived into clicking on the link that contained the latest information about COVID-19. Despite various ways proposed to overcome this problem, phishing attacks continue to increase. The focus of this study was Phishing Detection for Covid-19 Theme-Based Email and Weblinks using Machine Learning. The study was comprised of two parts. Web Links and Email Themed. Two types of datasets were selected for experiments. Dataset 1 contains Web URL data and was downloaded from Kaggle. Dataset 2 contains Email images and was downloaded from Google, and Bing search Engines. Different features were selected for the detection of Phishing. Python libraries and coding was used for the analysis. The voting technique of the Ensemble model was used. It was revealed during the study that Dataset 2 achieves the highest accuracy while Dataset 1 performs better for other performance measures. Interesting concepts were found during the study.

Keywords: Phishing; Email; URL; HTTP; DNS; ML

Introduction

Phishing is a fraudulent attack where an attacker attempts to get sensitive information from a victim via email, text messages, or websites. A type of social engineering attack is a Phishing email through which an attacker steals the victim's sensitive data e.g. bank credentials, health report, and home address. In just one week in April 2020, Google's AI-powered protection filter stopped over 18 million phishing and malware assaults with a Covid-19 theme. The global spread of the infectious illness covid-19 culminated in a pandemic that put millions of lives at risk. Working from home has become the new norm since Covid-19 broke out. COVID-related phishing emails are evolving to address the pandemic, and they are identified as donations to fake charities, malware delivery, or credentials. Credential theft and phishing account for 67% of all breaches before the pandemic and 22% of all breaches in 2020 are caused by phishing. Unscrupulous cybercriminals took advantage of the unawareness of the pandemic to make money. Phishing links are more likely to be clicked by most users. As a result, the phishing emails used "COVID" or "coronavirus" as subject lines to lure their victims into clicking on them. Electronic media has been critical in spreading information regarding the virus and its impact via a variety of mechanisms, including the continual transmission of local and worldwide updates, as well as the issuance of warnings and recommendations for coping with the virus and its aftermath. This pandemic provided additional possibilities for internet criminals to defraud people. Covid-19-related domains have seen a substantial surge in popularity as a result of people's curiosity in determining the threat's scope and identifying protective measures [1,2].

To mitigate the threat of phishing and spam various existing solutions use NLP or ML techniques [10]. ML algorithms such as RF and DT are utilized for the detection of spear-phishing emails.

Since phishing attacks have grown more complex and have changed their ways to defeat anti-phishing techniques, they continue to be a serious problem. Websites or emails that are phishing consist of fake URLs that look similar to popular and legal websites. Despite having distinct Uniform Resource Locators (URLs), fake websites have similar user interfaces. A user can identify a fake website by checking the URLs carefully. As a result, existing anti-phishing techniques, such as content-based and keyword approaches, are not able to stop phishing. Lack of research on the Ensemble model for the detection of Phishing on Covid-19 Themed Email or Covid19 themed URLs or Both together. As a solution to the above issue, we proposed a Machine Learning model to predict phishing using covid-19 themed emails and Web Links. The objective of this research is:

- Predict phishing using machine-learning algorithms for email content.
- Predict phishing using machine learning algorithms for web Links.
- Create an Ensemble machine-learning model for phishing prediction using ML Algorithms for email content and URL both.
- Evaluate the Performance of the Model in terms of Accuracy, Precision, Recall, and F1 score.

Related work

Many ways are used by phishers to believe victims that the link is legitimate as worked by Nurul Ainatasha Afandi et.al (2021) a Covid-19-based Hyperlink based phishing detection using the KNN algorithm was proposed. Two datasets were downloaded dataset 1 was taken from SpyCloud and Phishtank while Dataset 2 was taken from Kaggle and DomainTools. Each dataset contains 250 Phishing URLs and 250 Legitimate URLs. To evaluate the performance of the Hyperlink-based model a KNN algorithm was used. Domain Name, Generic_TLD, URL_Length, and Prefix were considered as features for the detection. Their model achieves an Accuracy rate of 97.80% and 99.60% respectively [4].

Phishing emails and emails claiming to provide information regarding the spread of the disease were the most prevalent types of unsolicited emails as discussed by Naci Akdemir., *et al*. (2021). By examining the tactics used by online offenders to leverage internet users' suspicions of the coronavirus through phishing emails, this research seeks to improve user protection. 208 Covid-19 phishing emails were subjected to content analysis. Phishers have created 9 different categories of email messages [5].

The majority of recent cybercrimes have been committed by malicious intruders, who are also to blame for an increasing range of cyber threats, such as thefts of identities and intellectual property, financial crimes, and attacks on vital infrastructure as worked by Jamil Ispahany et.al (2021) the authors proposed ML classification technique to detect the number of rising spiteful URLs during Covid-19 time. By using 5 features the model obtained an accuracy of 99.2%. The domain names were taken from WhoisDS and Domain Tools. According to their finding, the best results were generated by the KNN algorithm [6].

In addition to the COVID-19 pandemic, there was a worldwide scam and fraud epidemic that was just as dangerous as explained by Ali F. Al-Qahtani et.al (2021) their analysis identifies the main characteristics and techniques to preserve from attacks during Covid-19. To achieve their aim survey was performed. Only 54 studies and many reports have been identified to investigate those attacks [7].

Fake websites that look like real ones are created by phishers to steal sensitive information by sending emails as discussed by Ishita Saha et.al (2020) a data-driven framework presented by authors to detect phishing webpages by using deep learning. The dataset contains 10 attributes and was taken from Kaggle. Their model attains a 95% accuracy rate [8].

Despite the existence of numerous successful phishing email detectors, phishing emails continue to cost businesses and individuals millions of dollars each year studied by Gal Egozi., *et al*. (2018) proposed a phishing email detector using 26 features. According to their findings, the dictator was able to identify 95% of harmful emails and 80% of phishing emails. NLP techniques such as word count, stop word count, punctuation and uniqueness were the chosen features [9].

Materials and Methods

The main objective of this study is to create an Ensemble Machine Learning model to predict phishing using covid-19 themed emails and Web Links. Python Programming Language was used

to create the model and evaluate the performance of the model in terms of Accuracy, Precision, Recall, and f1 score. Our objective was achieved by following steps:

Step-1: Two types of the dataset were taken Dataset 1 was for Web Links and was collected from Kaggle, while Dataset 2 was for Email themed and was collected from Google, and Bing search Engines.

Step-2: The dataset was cleaned by removing duplicates, and irrelevant Data, Handle missing data.

Step-3: For the web link, we consider 3- features Address Bar, Length of URL, and Redirecting. For Email Themed we consider 3-features Sender's address, grammatical mistakes, and the Signature of the Sender **Step-4:** The model was created and trained using Python coding.

Step-5: Finally, the performance of the Ensemble Model was evaluated by applying 5-Ml Algorithms.

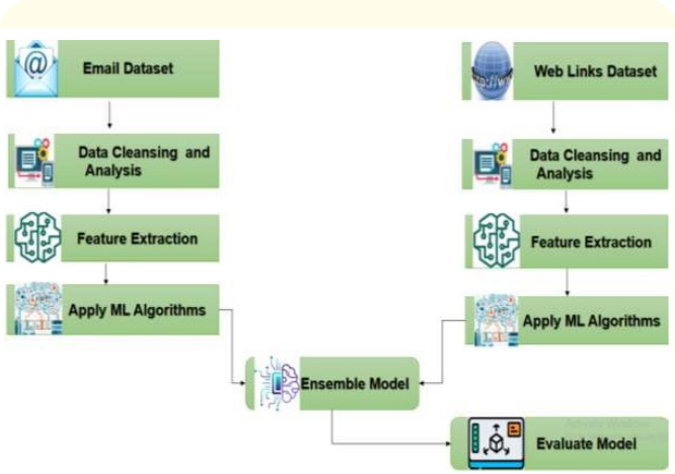

Figure 1: Research methodology.

Results and Discussion

The purpose of this study is to propose an Ensemble ML model to predict Phishing using Covid-19 web links and Email themes. This study was performed in two parts:

- Web Link Themed
- Email Themed

We have constructed two types of Datasets for our study. Dataset 1 for Web Links and Dataset 2 for Emails. Dataset 1 was collected from Kaggle and Dataset 2 was collected from the Search engine i.e. Google. As defined in below table 1.

Table 1: Web Links Dataset.

Dataset	Sources	Phishing	Legitimate	Features
Dataset 1	Kaggle	138	362	Text, Numeric
Dataset 2	Google	30	70	Images

Table 1 depicts the description of the Dataset. Dataset 1 [11] contains a total of 4530 data after preprocessing 43 were phishing dataset and 855 was Real dataset generated while dataset 2 contain a total of 200 email images after pre-processing we have 100 Phishing images and 100 Real images. The following features were selected for the detection of phishing URLs as shown in Table 2.

Dataset	Key Features	Description
Dataset 1	Address Bar	Length and detail about IP addresses
	Length of URL	Length of Characters
	Redirecting	By using // to redirect the page
Dataset 2	Grammar of Text	Grammatical mistake in the main Text Body
	Sender's Email	The email of the Sender i.e. domain Name, @, Numeric, etc.
	Signature of Sender	

Table 2: Description of Key Features.

The Table 2 Describes the Key features of our Dataset. Dataset 1 contains the length of the URL and the Address Bar which includes the number of slashes, the Number of hyphens, and the Length of the Domain name and contains dots. Dataset 2 has features such as Grammatical Mistakes, Sender's Email, the signature, etc.

Part 1: Web Link-themed detection

Phishing domains, also called fraudulent domains, are URL schemes that appear suspicious for a variety of reasons**.** The following are the features for the detection of Fake and Real links.

Address bar

Someone can be trying to steal your personal information using the IP address instead of using Domain Name. The URL http://125.98.3.123/fake.html is an example of a fake URL. It is even possible to use hexadecimal code to convert IP address as depicted below in the link "http://0x60.0xDC.0xCA.0x52/7/paypal.ce/index.html". The address bar contains @, Domain Name, depth of URL, prefix, etc.

Rule 1

IF {the address bar contains part of an IP address then it is Phishing else it is legitimate.

Length of URL

The average URL length was calculated using the dataset. In the study, phishing URLs were classified as such if the URL length exceeded 54 characters. We have found 250 URLs' length is less than 52 while 250 URLs are equal to or greater than 52 characters.

Rule 2

IF {Length of URL characters ≥ 52 and ≤ 75 URL is Phishing, else length of URL is < 52 then URL is Legitimate URL}.

Redirecting

If the URL path contains "//", the user will be redirected to another website. For example "http://www.legitimate.com/http://www.phishing.com". The "//" is examined at the location where it appears. If the URL begins with "HTTP", then the "//" should be placed in the sixth position. Alternatively, if the URL uses "HTTPS", then the "//" should appear seventh.

Rule 3

If {if the last existence of // is > 7 then the URL is Phishing, else URL is Legitimate}. Table 3 Frequency of Most Frequently used words.

Most Frequent words	Number of Occurrence
.	1642
https	2087
Coronavirus	890
/	900
-	1524
Covid	536

Table 3: Describes the frequency of the most frequently used word in both datasets. It can be observed that (.) has the most frequent to be use.

Evaluation of Algorithms

We have two types of datasets and each dataset has different features. 80% of the dataset was used for training and 20% for testing.

Accuracy

A measure of the proportion of appropriately forecasted observations to the total number of observations.

TN

A=TP++FN+FP+TN -- (a)

TP

Precision

A measure of the proportion of observations predicted correctly to be positive compared with all other observations. TP *A*=+FP-- (b) TP

Recall

In a given class, recall refers to how many positive observations were correctly predicted.

TP *A*=+FN-- (c) TP

True positive (TP)

Data points are classified as positive by the computer based on their actual effects.

False positive (FP)

Data points that were mistakenly interpreted as positive by the algorithm had negative effects in reality.

Table 4 describes the performance of 5-algorithms and the Ensemble Model in terms of Accuracy, Precision, Recall, and F1 score. The ensemble Model was used to create multiple models and combine them to provide the final prediction. Voting Techniques of the Ensemble Model were applied. It can be observed that the majority of algorithms achieved the highest Accuracy. The Overall Accuracy and Precision of the Ensemble Model was 0.79 while the Ensemble model attained the highest Recall.

Part 2: Email-themed detection

A method of transmitting and receiving messages using electronic devices is electronic mail (email). Email is also one of the ways to steal confidential data. We have collected 100 images from Google, and Bing search Engines. After collecting images, text extraction was performed using Python libraries [3].

ML Algorithms	Accuracy	Precision	Recall	F1 score
Naïv e Bayes	0.84	0.86	0.92	0.69
K-Nearest	0.78	0.83	0.89	0.84
Random Forest	0.84	0.76	0.78	0.69
Logistic Regression	0.84	0.86	0.92	0.79
Decision Tree	0.84	0.86	0.92	0.89
Ensemble Model	0.79	0.79	0.96	0.86

Table 4: Performance evaluation of Ensemble Model for Dataset 1.

ML Algorithms	Accuracy	Precision	Recall	F1 score
Naïv e Bayes	0.76	0.75	0.78	0.81
K-Nearest	0.73	0.75	0.78	0.78
Decision Tree	0.88	0.93	0.93	0.93
Random Forest	0.94	0.91	0.92	0.91
Logistic Regression	0.97	0.72	0.80	0.68
Ensemble Model	0.87	0.65	0.66	0.70

Table 5: Performance evaluation of algorithms for Dataset 2.

The performance of the 5-algorithm in terms of Accuracy, Precision, Recall, and F1 score and Ensemble Model is discussed in table 5. It can be observed that Logistic Regression achieves the highest Accuracy and the Decision Tree attained the highest f1 score. It can be said that Naïve Bayes works average.

Comparison of Performance Dataset1 and Dataset 2

Figure 2 depicts the overall performance of the Ensemble Model over Dataset 1 and Dataset 2. We can say that Dataset 2 attains good Accuracy while Dataset 2 works better than Dataset 1 in terms of other performance measures.

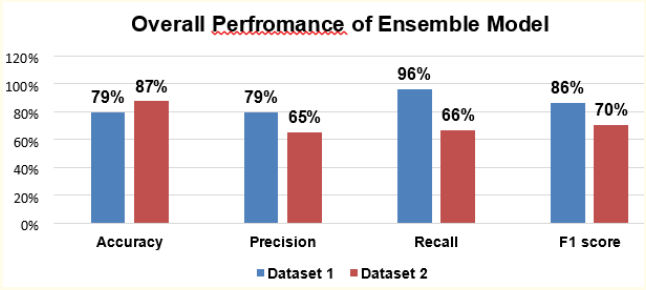

Figure 2: Overall Performance of Ensemble Model.

Conclusion

The threat of phishing URLs in cyber security can steal sensitive information from users. URLs can be sent via email. Phishers design phishing URLs in various ways to bypass detection techniques. Therefore, the objective of this study is to create an Ensemble detection model for phishing URLs and emails. Two types of datasets were selected Dataset 1 contains a description of URLs attack and was downloaded from Kaggle while Dataset 2 contains email images downloaded from Google, and Bing search engines. Python programming was used for pre-processing and evaluation of algorithms. 5-well known ML algorithms (Naïv e Bayes, SVC, K-Nearest, Random forest, and Logistic Regression) were chosen. The accuracy rates for Dataset 1 and Dataset 2 were 79% and 81% respectively based on six features. It was observed Dataset 1 attained the highest Precision, Recall, and F1 score. It can be concluded that Ensemble Model works better with Dataset 1. COVID-19's phishing detection model offers a promising solution for reducing COVID-related phishing URLs as well as emails. In light of the results, future research should focus on understanding attackers' behavior and profiling their methods and evaluating the performance of more features using other ML algorithms.

Bibliography

1. Clark JW. "Trends in social engineering: Securing the weakest link". NSI.

2. Kumaran N and Lugani S. "Identity and security. Protecting businesses against cyber threats during COVID-19 and beyond".

3. Dewis Molly and Thiago Viana. "Phish Responder: A Hybrid Machine Learning Approach to Detect Phishing and Spam Emails". *Applied System Innovation* 5.73 (2022): 2-19.

4. Afandi Nurul A and Isredza R A Hamid. "Covid-19 Phishing Detection Based on Hyperlink Using KNearest Neighbor (KNN) Algorithm". *Applied Information Technology and Computer Science* 2.2 (2021): 387-301.

5. Akdemir Naci and Serkan Yenal. "How Phishers Exploit the Coronavirus Pandemic: A Content Analysis of COVID-19 Themed Phishing Emails". (2021).

6. Ispahany Jamil and Rafiqul Islam. "Detecting Malicious Urls of COVID-19 Pandemic Using ML Techniques". 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops), (2021).

7. Al-Qahtani Ali F and Stefano Cresci. "The COVID-19 Scamdemic: A Survey of Phishing Attacks and Their Countermeasures during COVID-19". *IET Information Security* (2022).

8. Saha Ishita., *et al*. "Phishing Attacks Detection Using Deep Learning Approach". Proceedings of the Third International Conference on Smart Systems and Inventive Technology (2020).

9. Egozi Gal and Rakesh Verma. "Phishing Email Detection Using Robust NLP Techniques". 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (2018).

10. Abdelhamid Neda., *et al*. "Phishing Detection: A Recent Intelligent Machine Learning Comparison Based on Models Content and Features". 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), (2017).

11. Vrbanciˇcˇ Grega., *et al*. "Datasets for Phishing Websites Detection". Elsevier (2020).

12. Kawaoka Ryo., *et al*. "A First Look at COVID-19 Domain Names: Origin and Implications". In Proceedings of the Passive and Active Measurement Conference 2021 (PAM 2021), (2021).

13. Aljofey Ali., *et al*. "An Effective Detection Approach for Phishing Websites Using URL and HTML Features". *Scientific Reports* (2022).