Identification of Phishing Website Using Machine Learning

Kiran D. Tandale, MGM's Jawaharlal Nehru Engineering College, Aurangabad, India, kirantandale592@gmail.com

Dr. Sunil N Pawar, Associate Professor, MGM's Jawaharlal Nehru Engineering College, Aurangabad, India, sunilpawar@jnec.ac.in

Abstract **Phishing is a branch of information security through which attackers can gain access to sensitive user credentials by using counterfeit websites closely resembling legitimate websites. Phishing attacks are the most common form of cyber-attacks achieved by cleverly disguising website URLs to trick credulous users. With increasing number of new phishing attacks, the use of machine learning algorithms to classify websites as phishing and legitimate has been proposed in this paper. The dataset for this study comprises of 96,018 URLs comprising of both phishing and legitimate websites. The URLs have been parsed using Pandas and Urllib in order to extract useful features that could help in phishing detection. Different ML algorithms such as Random Forest, Decision Tree, Adaboost, Fuzzy Pattern Trees etc. have been implemented on the data and a comparison is drawn between them. Random Forest Algorithm proved to be the most accurate algorithm with 95.82% accuracy.**

Keywords —Phishing, Website, Machine Learning, URL

I. Introduction

Phishing is fraudulent activity which involves the use of counterfeit websites by attackers to steal personal and sensitive user details. These may involve email login credentials, one-time password for transactions, bank account username and password, credit & debit card pin numbers and so on. In Phishing, the attacker appears to be a reputable entity and tricks the user into sharing sensitive details. Phishing involves tricking the user to share details with the attacker which makes it a simpler way of breaking into a computer's defense system in comparison to hacking. Phishing attacks are often carried out through e-mails containing spoofed logos with malicious links which appear to be legitimate to an unsuspecting user. Based on data [1], a new phishing website is created every 20 seconds on the internet. Also, recipients open 70% of the phishing attempts they receive. From [2], 0.47% of bank account holders become targets of phishing attacks each year leading to $2.4M to $9.4M losses per million clients. These statistics reveal the ease with which attackers can target unsuspecting users and the need to have a robust phishing attack detection mechanism.

The process of carrying out a phishing attack is as follows. The attacker mimics the login page of a popular website and registers it with a URL which looks very similar to a legitimate website. An e- mail is then sent to the user with the link of the phishing website. The body of the e-mail is disguised to make it seem legitimate to the person reading it. The user then clicks on the link and enters the login credentials. The login page of the cloned website has a script running at the backend which extracts the credentials entered by the unsuspecting user and makes it available to the attacker. The attacker can then utilize these credentials in the legitimate website and exploit the user. This process is illustrated in Fig. 1.

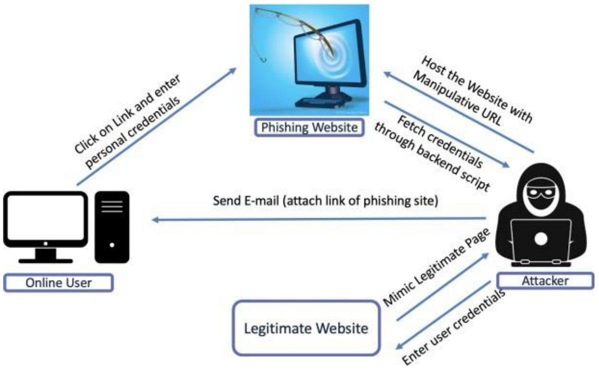

Phishing Website
Host the Website with Manipulative URL
Fetch credentials through backend script
Click on Link and enter personal credentials
Send E-mail (attach link of phishing site)
Online User
Attacker
Mimic Legitimate Page
Legitimate Website
Enter user credentials

Fig. 1. Phishing Mechanism

The commonality of all phishing attacks is the disguise of the website URL. The victims of phishing attacks are most often tricked by the URL of the phishing website. There are two methods used by attackers such that Cyber squatting and Typo squatting. Cybersquatting is the process of URL hijacking. The attacker buys the domain name of an already established company which does not have a website related to the domain name. Typo squatting refers to buying a website URL similar to a legitimate website but containing a typographical error. An example of this is google.com and goggle.com. Often, internet users make typing errors while entering the website URL which is exploited by attackers.

Besides these, the attacker may also choose to manipulate the URL by altering the sub- domain names, query lengths, adding redirect requests or making the URL excessively long. Since phishing data is easily available in phishing databases such as Phish tank and Open Phish, once a website is suspected of being related to phishing, the attacker can easily modify the website URL by altering the sub-domain names to make a new website. Therefore, there is a need for an intelligent method for identifying phishing URLs and reduce phishing attacks. Data mining techniques can help in classification of website URLs into phishing and legitimate URLs.

In this study, 7 machine learning algorithms have been tested on the dataset. The results of the different algorithms have been tabulated below.

The accuracy of different algorithms was determined for the original feature set as well as for PCA applied feature set as depicted in Table I. The accuracy of Random Forest algorithm was identified as the best algorithm with 95.82% after PCA[3]. Logistic regression was the worst performing algorithm amongst all of models[5]. Decision Tree, Gradient Boosting, Fuzzy Pattern Classifier and Adaboost gave accuracies which were close to Random Forest Algorithm[8][9].

.

Table- I: Accuracy of different algorithms

Accuracy of ML Algorithms		
Algorithm	**Independent Accuracy (%)**	**Accuracy with PCA (%)**
Random Forest	95.33	95.82
Decision Tree	94.09	94.26
Gradient Boosting	92.19	92.22
Fuzzy Pattern Tree	91.22	92.23
Adaboost	91.00	90.64
Gaussian NB	83.28	85.17
Logistic Regression	73.78	82.89

A. *Regression Algorithm*

Machine learning algorithms can also be divided as parametric learning model and nonparametric learning model. Algorithms that have strong assumptions in the learning process and that simplify the function to the known form are known as parametric machine learning algorithms. Linear regression and logistic regression are the examples of parametric machine learning algorithms. Regression algorithms deal with modeling the relationship between variables that are refined iteratively using a measure of error in the predictions made by the model. Linear regression is an approach to model the relationship between a scalar-dependent variable y and one or more explanatory variables (or independent variables) denoted x. Linear and logistic regressions are the major algorithms in predictive modeling. And for this project regression algorithm is used.

The organization of the remainder of the paper is as follows. Section II defines the methodology adopted in this paper to classify websites into phishing and legitimate. Section III describes the implementation. Section IV describes the results of the work. Section V contain concluding remarks of paper.

II. METHODOLOGY

Determining whether a given website URL is phishing or legitimate is a binary classification problem which can be solved with the help of labelled data on which supervised learning can be applied. Data collection for this problem requires recent website URLs belonging to both classes - phishing and legitimate. This is followed by preparation of the dataset by extracting relevant features which helps in distinguishing phishing websites from legitimate websites. The features need to be processed in order to give as input to the machine learning algorithm. Then the model is trained using the training set and its accuracy is determined on the testing set. The flow chart depicting the methodology is summarized in the Fig. 2.

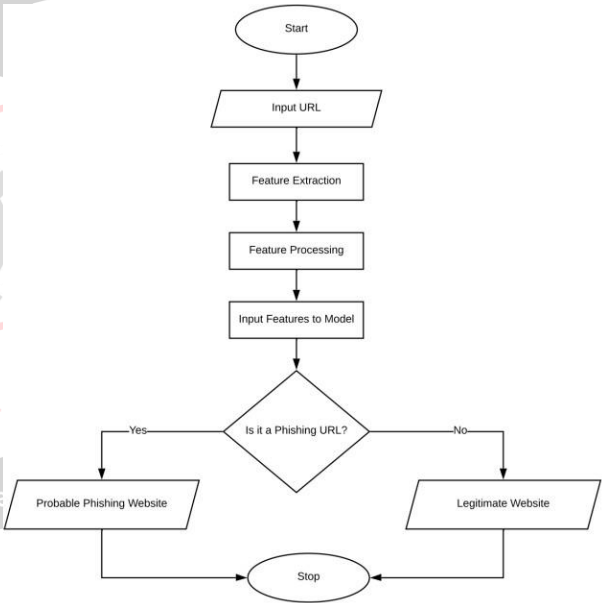


Fig. 2. Methodology

A. *FEATURE SELECTION*

1) *URL Based Features:*

- IP Address: The use of IP address or hexadecimal characters in the domain of the URL instead of a textual domain name can be a probable phishing website. An example of IP address in URL is http://102.24.134.12/page.html. In 46.66% of cases, the use of IP address or hexadecimal characters has been linked to phishing websites and suspicious activity according to a study in [17].
- @ symbol in URL: The use of '@' in the URL causes the browser to disregard all the contents prior to the symbol. Often, the phishing website address follows the '@' symbol. This is often

DOI : 10.35291/2454-9150.2021.0025

utilized as a means to exploit phishing since users do not often read the full URL. The appearance of '@' has been found in 20% of the phishing websites in the study done in [17].

- HTTPS in the middle of URL: The presence of "https" in the domain of the URL is used by phishers for tricking people. https.paypal.com-account-update.e3d3idw3k4security-alert.cenksen.com.tr/paypal.com/ is an example of how users are deceived by http in URL domain.

2) *Domain Based Features:*

- Page Rank of the Website: The importance of a website is often marked by its page rank. Alexa maintains a database of websites is used to determine the page rank of a given website. A phishing website is usually unranked or very lowly ranked in comparison to legitimate websites. This is again an important differentiating factor between phishing and non-phishing pages.

- Age of the Domain: A phishing website has a very short lifetime in comparison to legitimate websites which are usually up for a very long duration of time. The Whois API helps to determine the age of the domain which is a useful metric for differentiating between phishing and non-phishing pages.

- Validity of the Website: The Google Whois API helps to identify whether the website is still operational or not. Most of the phishing sites have a short lifetime and are pulled down once detected of suspicious activity. The validity is an important marker that separates legitimate website from phishing websites.

III. Implementation

The dataset for classification of phishing and legitimate websites has been prepared based on the set of URLs from [18]. A total of 96,018 URLs have been taken into consideration with 48,009 phishing and legitimate URLs each. The phishing URLs have been taken from PhishTank database which provides valid and suspected phishing URLs. The legitimate websites have been taken from Open Directory Project (DMOZ).

The first step in processing the database was to determine the useful features to give as input to the machine learning algorithm. Dataset exploration was carried out using Python and attributes for classifying URLs was determined. The features selected for Phishing URL Detection are described in Section IV. Feature extraction from the URLs was done using a python library called "urlib" which parsed the URL string and split it into network protocol, domain names, sub-domain names and query strings.

The dataset in our study was split into training and testing set in the ratio 80:20. The training set with the extracted features were given as input to different machine learning classification algorithms.

The accuracy of classification along with precision, recall and F1 Score was determined on the examples in the testing set. The results were then compared and analyzed for the different models. Finally, the model was tested on real world phishing examples to determine its robustness.

Based on Figure 3, admin can filter which URLs are blacklisted and which are not blacklisted by copy and pasting the URLs at "Site" row.

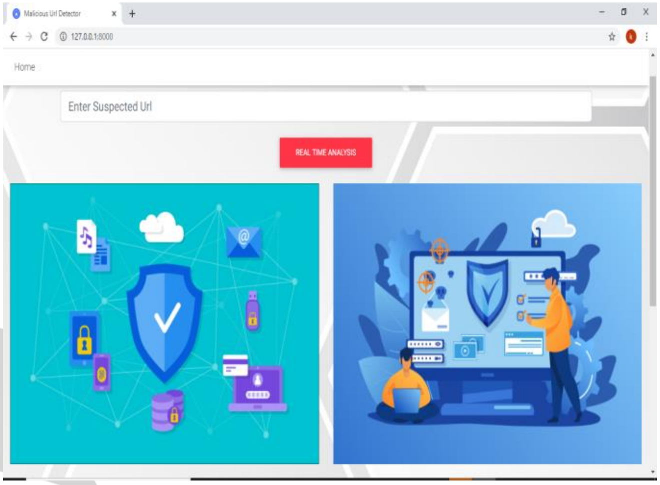


Fig. 3 Project Admin Panel

A. *Sentiment Analysis*

It is natural language processing technique which is used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

1) *Types of sentiment Analysis*

If polarity precision is important to your business, you might consider expanding your polarity categories to include:

- Very Positive
- Positive
- Neutral
- Negative
- Very Negative

This is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:

- Very Positive- five star
- Very Negative- One star

a) Emotion Detection

This type of sentiment analysis aims to detect emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithm.

b) Aspect-Based Sentiment Analysis

Usually, when analyzing sentiments of texts, let's say product reviews, you'll want to know which particular aspects or features people are mentioning in a positive, neutral, or negative way. That's where aspect-based

DOI : 10.35291/2454-9150.2021.0025

sentiment analysis can help, for example in this text: "The battery life of this camera is too short", an aspect-based classifier would be able to determine that the sentence expresses a negative opinion about the feature battery life.

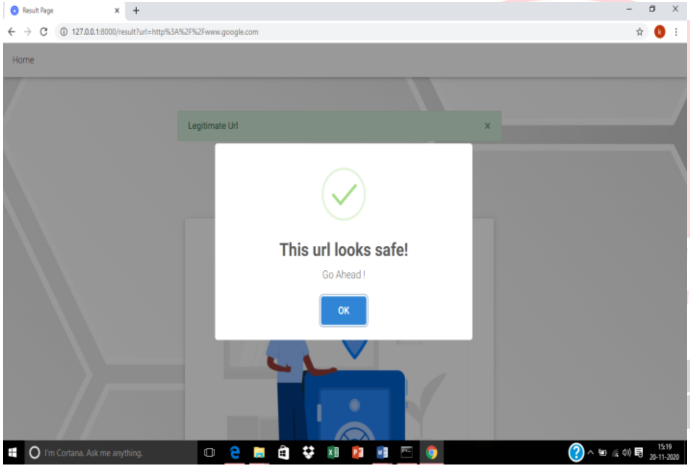
c) Multilingual sentiment analysis

Multilingual sentiment analysis can be difficult. It involves a lot of preprocessing and resources. Most of these resources are available online (e.g. sentiment lexicons), while others need to be created (e.g. translated corpora or noise detection algorithms)

IV. RESULTS AND DISCUSSION

In this project, various website was tested and found good accuracy. Here two different website's results are attached. First result is for www.google.com which showing that the website looks safe (fig. 4) and showing more information about the same website where it showing various parameters like organization name, address, email and domain name (fig5). The second result is of ord-amazsn.com which showing given website is malicious (fig 6) and showing more information about the same website (fig 7).

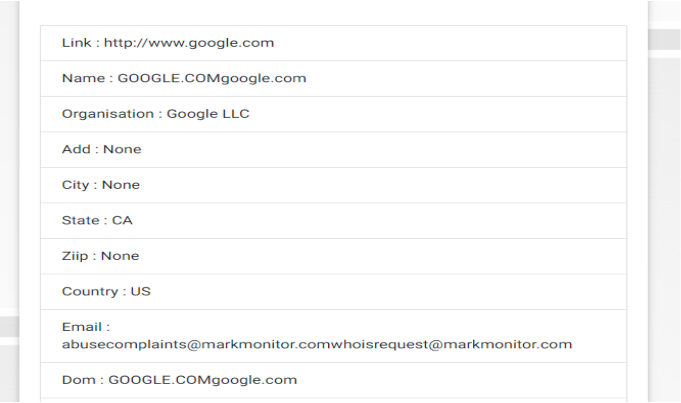
The actual visualization of results pages are as follow:


Result Page
127.0.0.1:8000/result?url=http%3A%2F%2Fwww.google.com
Home
Legitimate Url
OK
I'm Cortana. Ask me anything.
15:19
20-11-2020

This url looks safe!

Go Ahead !

Fig. 4 Legitimate website Result



Link : http://www.google.com

Name : GOOGLE.COMgoogle.com

Organisation : Google LLC

Add : None

City : None

State : CA

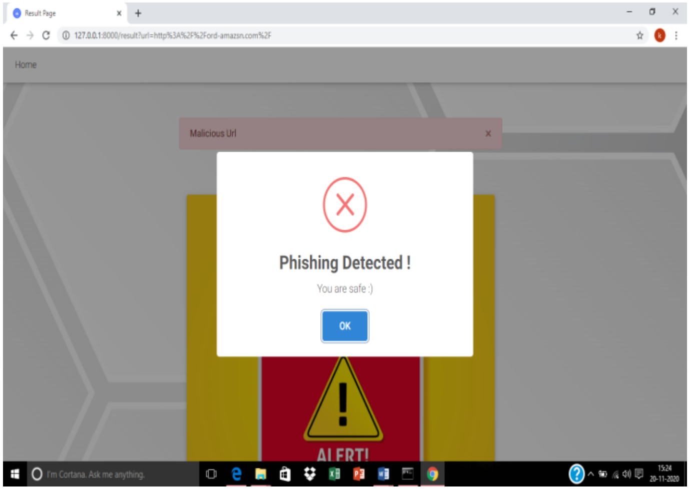
Ziip : None

Country : US

Email : abusecomplaints@markmonitor.comwhoisrequest@markmonitor.com

Dom : GOOGLE.COMgoogle.com

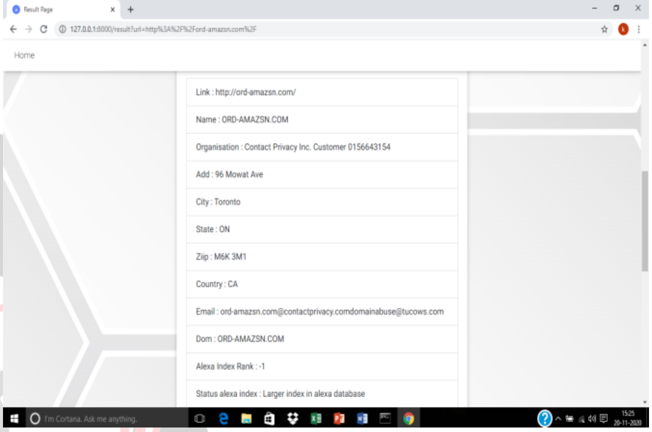
Fig. 5 Backend Reply for Legitimate website Result


Result Page
127.0.0.1:8000/result?url=http%3A%2F%2Ford-amazsn.com%2F
Home
Malicious Url
OK
ALERT!
I'm Cortana. Ask me anything.
15:24
20-11-2020

Phishing Detected !

You are safe :)

Fig. 6 Malicious website Result


Result Page
127.0.0.1:8000/result?url=http%3A%2F%2Ford-amazsn.com%2F
Home
I'm Cortana. Ask me anything.
15:25
20-11-2020

Link : http://ord-amazsn.com/

Name : ORD-AMAZSN.COM

Organisation : Contact Privacy Inc. Customer 0156643154

Add : 96 Mowat Ave

City : Toronto

State : ON

Ziip : M6K 3M1

Country : CA

Email : ord-amazsn.com@contactprivacy.comdomainabuse@tucows.com

Dom : ORD-AMAZSN.COM

Alexa Index Rank : -1

Status alexa index : Larger index in alexa database

Fig. 7 Backend Reply for Malicious website Result

V. CONCLUSION

This paper explains the existing security problems in today's digital world with respect to phishing and the process through which phishing is carried out. Phishing is a serious security concern which may lead to loss of sensitive personal information due to clever disguising of phishing mails by attackers. This work mainly focuses on identifying features useful for detecting phishing websites based on solely the URL of the website and applying machine learning algorithms to classify websites into legitimate and phishing. The study involves comparison of results of 7 machine learning algorithms with Random Forest algorithm emerging as the most accurate and hence, most suited algorithm for this binary classification

REFERENCES

[1] https://www.merchantfraudjournal.com/phishing-attack-statistics-2019/

[2] https://www.businesswire.com/news/home/20091202005153/en

[3] J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology,", 5th International

DOI : 10.35291/2454-9150.2021.0025

Conference on Signal Processing and Integrated Networks (SPIN), 2019 pp. 425–430, 2018.

[4] Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[5] T. Peng, I. Harris, and Y. Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning,” Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[6] M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018– Janua, pp. 1–5, 2018.

[7] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. ICICCT, pp. 949–952.

[8] K. Shima et al., “Classification of URL bitstreams using bag of bytes,” in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[9] A. Vazhayil, R. Vinayakumar, and K. Soman, “Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks,” in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1– 6.

[10] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, “Boosting the Phishing Detection Performance by Semantic Analysis,” 2017, IEEE International Conference on Big Data (big data),pp. 1063-1070.

[11] A. Desai, J. Jatakia, R. Naik, and N. Raul, “Malicious web content detection using machine leaning,” RTEICT 2017 - 2nd IEEE International Conference on Recent Trends in Electronic Information Communication Technology Process, vol. 2018–Janua, pp. 1432–1436, 2018.

[12] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, “A novel approach for phishing detection using URL-based heuristic,” 2014 International Conference on Computing, Management, Telecommunication, ComManTel 2014, pp. 298–303, 2014.

[13] S. Marchal, J. Francois, R. State, and T. Engel, “PhishScore: Hacking phishers’ minds,” Proc. 10th Int. Conf. Netw. Serv. Manag. CNSM 2014, pp. 46–54, 2015.

[14] A. A. Ahmed and N. A. Abdullah, “Real time detection of phishing websites,” 7th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEEE IEMCON 2016, 2016.

[15] Abu-Nimeh, Saeed & Nappa, Dario & Wang, Xinlei & Nair, Suku. A comparison of machine learning techniques for phishing detection. ACM International Conference Proceeding Series. 2007.

[16] Jalal, Kahksha & Naaz, Sameena. Detection of phishing website using machine learning approach, International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019).

[17] Maher Aburrous, Hossain, M.A., KeshavDahal and FadiThabtah, “Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies”, Cognitive Computing, DOI 10.1007/s12559-010-9042-7, Vol. 2, pp. 242-253, 2010.

[18] S. Marchal, J. Francois, R. State, and T. Engel. PhishStorm: Detecting Phishing with Streaming Analytics. IEEE Transactions on Network and Service Management (TNSM), 11(4):458-471, 2014.