Web Phishing Detection using Machine Learning

Article *in* International Journal of Innovative Technology and Exploring Engineering · March 2022

DOI: 10.35940/ijitee.C9804.0311422

Natarajan Kumaran
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University
15 PUBLICATIONS **41** CITATIONS

Web Phishing Detection using Machine Learning

N Kumaran, Purandhar Sri Sai, Lokesh Manikanta

***Abstract:** A web service is one of the most important Internet communications software services. Using fraudulent methods to get personal information is becoming increasingly widespread these days. However, it makes our lives easier, it leads to numerous security vulnerabilities to the Internet's private structure. Web phishing is just one of the many security risks that web services face. Phishing assaults are usually detected by experienced users however, security is a primary concern for system users who are unaware of such situations. Phishing is the act of portraying malicious web runners as genuine web runners to obtain sensitive information from the end-user. Phishing is currently regarded as one of the most dangerous threats to web security. Vicious Web sites significantly encourage Internet criminal activity and inhibit the growth of Web services. As a result, there has been a tremendous push to build a comprehensive solution to prevent users from accessing such websites. We suggest a literacy-based strategy to categorize Web sites into three categories: benign, spam, and malicious. Our technology merely examines the Uniform Resource Locator (URL) itself, not the content of Web pages. As a result, it removes run-time stillness and the risk of drug users being exposed to cyber surfer-based vulnerabilities. When compared to a blacklisting service, our approach performs better on generality and content since it uses learning techniques.*

***Keywords**: Security; Web Services; URL; Vulnerabilities*

I. INTRODUCTION

We do the majority of our work during the day on internet platforms. Using a system and the broadband connection in a variety of ways makes our work and personal lives easier. It provides us to accomplish sales and activities in industries such as business, medical, academia, information, finance, aeronautics, exploration, infrastructure, enjoyment, and welfare programs promptly. With the advancement of mobile and wireless technology, drug addicts who need to pierce a unique network can now be readily accessible to the web any day at any time. Although this arrangement is extremely convenient, it has highlighted major information security bugs. As a result, now need to drug addicts in cyber to take precautions from computer security is revealed. Cybercriminals, rovers, or non-malicious (fair-limited) bushwhackers and data theft are all capable of carrying out attacks. The goal is to get to the system or the content it stores or to retrieve specific data from various methods. Bushwhackers seek to obtain a lot of information and or plutocrat by contacting a wide range of target druggies. As according Kaspersky's analysis, the estimated price of an attacker by 2019 is between $ 108K and $ 1.4 billion (depending on the scale of the attack). Furthermore, the plutocrat spent roughly $ 124 billion on worldwide security products and services. Phishing is the most common sort of cybersecurity assault, and bushwhackers are no exception. Phishing assaults are typically simple since most victims are unaware of the complexities of web operations, computer networks, and related technology, making them ideal prey for being misled or caricatured. It's far easier to phish unwitting drug addicts using fake websites and entice them to click on the websites in exchange for a prize or offer than it is to target the information security system. A vicious site is created to have a similar look and feel, and it looks to be authentic in appearance because it uses the association's ensigns and other copyrighted content. As a result of many drug addicts unknowingly accessing the phishing website URLs, the person and the organization involved suffering significant financial and reputational losses. The most common and dangerous of these types of cyberattacks was phishing. Cybercriminals typically use a dispatch or other social networking communication channels in this type of attack. Bushwhackers approach the drug addicts by claiming that the money was transmitted through a reputable site, other than banking, an e-commerce platform, and something related. As a result, someone attempts and try crucial information from them. Bushwhackers also use this information to puncture their victims' accounts. As a result, it results in financial loss and irreversible harm.

II. LITERATURE SURVEY

The current circumstance is that the population's maturity has been wisecracked, causing them to unknowingly give their private information to hackers. Several banned websites have already been established to seem like that of an actual point of contact through obtaining stoners' private information. Passcode, savings account, and shipping information are just a few examples. Late in 2016, the amount of hacking activities was at an all-time high since the company started monitoring this in 2004. The overall identified phishing attacks in 2016 were 1,609. This represents a 65 percent increase over 2015. Within the final quarter of 2004, there would be scamming attempts each month. Machine Learning was used to find the phishing website. The use of machine literacy to surround the supplied features is the basis of Grounded Malware Monitoring Systems. Features are generated by assembling items in a specific order, such as URLs, sphere names, website features, and website content.

Manuscript received on February 28, 2022.
Revised Manuscript received on March 27, 2022.
Manuscript published on March 30, 2022.
* Correspondence Author

N Kumaran, Assistant Professor, Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram (Tamil Nadu), India. Email: nkumaran@kanchiuniv.ac.in

Purandhar Sri Sai*, B.E Student, Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram (Tamil Nadu), India. Email: purandharsrisai@gmail.com

Lokesh Manikanta, B.E Student, Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram (Tamil Nadu), India. Email: lokeshmanikantapuvvadi@gmail.com

*Retrieval Number: 100.1/ijitee.C98040311422
DOI: 10.35940/ijitee.C9804.0311422
Journal Website: www.ijitee.org*

*Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.*

Because of its nonlinear system, it has a high level of fashion ability in terms of web security, particularly for the detection of anomalies on internet spots. The features retrieved utilizing machine literacy approaches are compared to extracting features through URLs, primary law, or third-party services. A process of machine trust ability on a particularity meant for the reflection of the besieged deceit of stoners through electronic communication is a relevant approach for detecting these attacks.

This method can be used to find phishing websites or textbook dispatches sent over email to confuse the victims. This method was presented by S. Marchal et al. to distinguish Malicious URLs based on the assessment of legitimate point garçon record data. By the off operation or the detection of a malicious site. Open source demonstrates several remarkable characteristics, including high proximity, total autonomy, excellent linguistic flexibility, quickness in choosing, inflexibility towards active phishing, and inflexibility towards development in phishing methods. Mustafa Aydin et al. presented the bracket method to fraudulent site detection that involves rooted websites 'URL properties and evaluating subset- grounded Point selection approaches. For the detection of phishing websites, it uses point birth and selecting styles.

Fadi Thabtah et al. evaluated vast numbers of ML methods to actual malware datasets and according to many parameters. The goal of this comparison is to highlight the benefits and drawbacks of ML predictive models, as well as their real performance in phishing attempts. Covering approach models are more appropriate as anti-phishing results, according to the experimental results. Muhemmet Baykara et al. developed the Anti Phishing Simulator, which gives data on the phishing discovery challenge as well as how to detect phishing emails. Only utilize the textbook of the e-mail as just a term to execute complicated word processing, according to the study's recommendations.

III. PROPOSED SYSTEM

The lexical features are based on the observation that many unauthorized sites' URLs differ from legal sites' URLs. We can capture the quality for categorization purposes by analyzing lexical features. To extract bag-of-words, we first separate the two elements of a URL: the hostname and the path. We've discovered that phishing sites like longer URLs, more scenarios, more commemoratives in the sphere and path, and longer tokens.

Phishing and malware websites could either pretend to be a harmless bone by including popular brand names as commemoratives instead of those in the alternate-position sphere. In the event of phishing and malware websites, the IP address may be used directly to mask the suspicious URL, which is extremely rare in benign cases. Phishing URLs are also designed to include several misleading word commemoratives. We look for these security-sensitive words and save the double value in our features. Violent spots have traditionally been less popular than benign bones. As a result, point fashion ability can be considered a significant factor. The use of host-based characteristics is based on the observation that aggressive spots are always found in lower-rated hosting hubs or areas. The dataset's URLs are all labeled, which is an advantage. We utilized the scikit-learn package to train two supervised literacy algorithms, arbitrary timber, and svm.

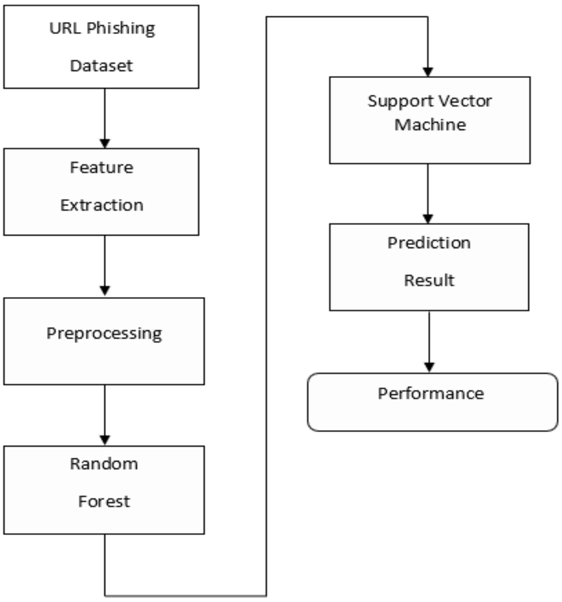

Fig 1. Architecture for proposed system

A. ALGORITHM

Step 1: Import the Dataset.
Step 2: Read the Dataset.
Step 3: Extract the data from the dataset for preprocessing.
Step 4: Make predictions for the test dataset.
Step 5: Applying Machine Learning algorithms to the dataset.
Step 6: Predict the best and worst accuracy algorithms from ML algorithms.

IV. METHODOLOGY

Modules:

- Data Collection
- Data Pre-Processing
- Feature Extraction
- Evaluation Model

A. Data Collection

The data for this project is a collection of records. This stage includes choosing a sample of all available information on which to work. Data, especially as the huge quantity of data whereby the target output has been established, is the starting point for machine learning challenges. Data that has been labeled contains information for which we have an answer.

B. Data Pre-Processing

Organize the data we've chosen by formatting, cleaning, and sampling it. Three common data pre-processing steps are:

- **Formatting:** That information we've chosen will not be in an easy-to-work-with format. The data could have been in a relational database which we'd like to export to a flat file, or it could have been in a unique file format that you'd like to export to a relational database or a text description.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

- **Cleaning:** Clearing information includes eliminating and replacing data that isn't present. There could be a situation when data is missing or imperfect, and we don't have all of the information we need to solve the problem. It is indeed likely that all these circumstances have to be removed. Moreover, a few of the characteristics might be sensitive data, which must be cleared or completely removed from the information.
- **Sampling:** There could be a lot more well-chosen data accessible than we need. Increased method execution durations and larger computational and storage requirements result from more information. We can choose a shorter sample size of the data sample before reviewing the complete dataset, which will allow us both to explore and develop ideas much faster.

C. Feature Extraction

The following stage is feature extraction, and that's an attribute extension that allows us to create more columns from URLs. Finally, we use a classifier algorithm to train our models. They take advantage of the obtained classified dataset. The remainder of our classified data would be used to validate the models. ML algorithms have been used to identify pre-processed data. That classifier utilized had been Random Forest.

D. Evaluation Model

The evaluation of a model is a key step in its development. It helps us to figure out which model perfectly describes the data but also how this might perform as in years ahead. To prevent overfitting, two very different methods require a test carried out to analyze the accuracy of the model. In evaluating the efficiency of each classification model, the median efficiency is employed. The final product will take the form that has been imagined. Information during classification is represented using graphical representations.

Accuracy is measured by the proportion of predictions made using the testing dataset. It's easy to calculate by dividing the total number of forecasts even by correctly predicted guesses. We calculate accuracy as the difference between actual and expected output calculate accuracy as

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Where TP = True Positives
TN = True Negatives
FN = False Negatives
FP = False Positives

V. PROJECT DESCRIPTION

That data contains several factors that should be considered when deciding whether a website URL is licit or phishing.

The factors for discovery and bracket of phishing websites are as follows

- Address Bar based Features
- Abnormal Based Features
- HTML and JavaScript Based Features
- Domain-Based Features

A. Address Bar based Features

1. Using the IP address
If the URL has an IP address rather than a sphere name, such as 125.96.2.121, a person can practically be assured that his private details are being stolen.

2. The Suspicious Part is hidden by a long URL
By selecting a long URL, phishers can hide the suspect portion of the URL inside the URL bar.

3. Applying URL shortening services The URL is very short
URL shortening is a mechanism on the Internet that allows a URL to be drastically reduced in length while still directing to the desired webpage.

4. URLs having @ symbol
When the @ sign appears in a URL, the cyber surfer ignores anything preceding the@ symbol, and the true address frequently appears after the@ symbol.

5. Redirecting using double slash
The presence of // in the URL route indicates that the stoner will be redirected to a different website.

6. Adding a domain prefix or suffix separated by (-)
The hyphen sign is hardly in use in licit URLs. To provide the idea that they'll be interacting with a reputable site, fraudsters frequently append prefixes or suffixes to the sphere name, separated by (-).

7. Multiple Subdomains and Subdomains
Let's imagine we're looking at the URL below. http//www.kanchi.ac.in/students/. The country-law top-position disciplines could be included in a sphere name.

8. HTTPs (Hyper Text Transfer Protocol with Secure Sockets Layer)
The existence of HTTPS is unquestionably important in determining the legitimacy of a website, but it is far from sufficient.

9. Domain Registration Length
Because a phishing website only exists for a short time, we feel that secure disciplines are paid for in advance numerous times. We discovered that the longest false disciplines were only utilized once in our sample.

10. Favicon
A favicon is a graphic image that is connected with a website.

11. Making use of a Non-Standard Port
That point is important for considering whether or not a specific service is over or unavailable on a specific server.

12. The presence of an HTTPS Token in the URL's Address part To mislead drug addicts, phishers may add the HTTPS commemorative to the spherical section of a URL.

B. Abnormal Based Features

1. Request URL Request
The URL test determines whether or not the external things featured within a website, such as photos, films, and sounds, are loaded from another location.

2. URL of Anchor
An anchor is a detail that is described by a tag. This attribute is referred to as the Request URL.

3. Links in Tags
Given that our research covers all perspectives that could be used inside the website supply code, Valid websites typically use tags to provide info regarding the actual Html page, as we've seen. To create a user script, use the script tags. Use link tags to find various internet resources. Those tags should all be related to a certain page of the site.

4. Server from Handler (SFH)
Because process should be performed upon that reported file, SFHs that contains an empty string or roughly: clean are suspicious.

5. Submitting to Email
An application form helps the customers can submit private details, which are then processed by a server. The customer's details could be sent to a phisher's private email.
6. Abnormal URL
That information may be obtained from the WHOIS network. A valid website's URL will generally include identification.

C. HTML and JavaScript Based Features

1. Redirect
This number of cases a webpage is being rerouted seems to be the distinguishing factor between phishing and legitimate websites.
2. Right-Click disablement
JavaScript is used by phishers to block the right-click on a feature, preventing customers from accessing and purchasing website programs is written. This function is used at the same time that Using on Mouseover to Cover the Link is handled.
3. Making use of Pop-Up Window
It is really rare as for come across a malicious website that asks visitors to provide private details through a pop-up window.
4. Redirection of the IFrame
An IFrame is a type of HTML tag that allows you as for embed another website inside the one you're now viewing.

D. Domain-Based Features

1. Age of Domain
The WHOIS network may be adjusted to extract this function. The bulk of such phishing websites is only operational for a small period. We can observe from the data also that the qualifying area must be at least 6 months old.
2. DNS Record
In the case of phishing websites, either the declared identity cannot be verified using the WHOIS database, or there are no facts to back up the claim. The website is rated as Phishing if a DNS document is blank sometimes no longer available; otherwise, this is categorized under Valid.
3. Web Traffic
The function calculates the number of visitors and the pages they visit to determine the overall reputation of a site.
4. Page Rank
PageRank is a market price that goes between 0 and 1. PageRank seeks to determine a website's popularity on the Internet.
5. Google Index
The function determines not whether a website should be indexed by Google. Whenever a domain is registered with Google, it shows up on the list.
6. The Number of Links Pointing to specific Page
Although a few connections are of the same size, the number of connections linking to a website shows its level of validity.

VI. RESULT

Machine learning methods were imported using the Scikit-learn library. Each classification is performed using a training set, and the performance of the classifiers is evaluated using a testing set. The accuracy score of classifiers was calculated to assess their performance.

VII. CONCLUSION

We discuss our large-scale system for automatically categorizing phishing runs in this design, which has a false positive rate with less than 0.1. In a fraction of the time, it takes a customized review procedure, our bracket system reviews millions of implicit phishing runner's responses. We reduce the amount of time that phishing runners can be active before we protect our druggies by automatically simplifying our blacklist with our classifier. Indeed, our blacklist strategy keeps us a step ahead of the phishers, thanks to a superb classifier and a robust system. Using the machine literacy method, we can only distinguish between phishing and legitimate URLs. In terms of the delicacy meter, this is what we obtained.

REFERENCES

1. Detecting Phishing Websites Using Machine Learning by Sagar Patil, Yogesh Shetye, Nilesh Shendage published in the year 2020.
2. Machine Learning-Based Phishing Attack Detection by Sohrab Hossain, Dhiman Sarma, Rana Joythi Chakma published in the year 2020.
3. Phishing website detection based on effective machine learning approach by Gururaj Harinahalli Lokesh published in the year 2020.
4. Research on Website Phishing Detection Based on LSTM RNN by Yang Su published in the year 2020.
5. Detecting Phishing Website Using Machine Learning by Mohammed Hazim Alkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen published in the year 2020.
6. Detection of Phishing Websites by Using Machine Learning-Based URL Analysis by Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri published in the year 2020.
7. Phishing Website Classification and Detection Using Machine Learning by Jitendra Kumar, A. Santhanavijayan, B. Janet, Balaji Rajendran, B.S. Bindhumadhava was published in the year 2020.

AUTHORS PROFILE

Dr N. Kumaran, received his Ph.D. from National Institute of Technology, Tiruchirappalli., M. Tech (Information Technology) from Sathyabama University, Chennai and B.E (Computer Science and Engineering) from Coimbatore Institute of Technology, Coimbatore-14, Tamil Nadu, India, in the year 2007 and 1998 respectively. He is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (Deemed to be University) Kanchipuram. He has served as a reviewer committee for an international journal. He has been an active member of professional bodies, such as IAENG and Judge of Toycathon 2021 (Ministry of Education and Innovative Cell). He also received best teacher award in the 2016 in his university. His areas of interest include Video Analysis, Machine Learning, Internet of Things and Computer Networks. He has got around 20 years of teaching experience various Institutions.

Purandhar Sri Sai is in his final year of Bachelor of Engineering in the department of Computer Science and Engineering (CSE) at Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Deemed to be University, Enathur, Kanchipuram, Tamil Nadu, India. His areas of interest include Full Stack Web Development, Machine Learning, and Data Science.

Lokesh Manikanta is in his final year of Bachelor of Engineering in the department of Computer Science and Engineering (CSE) at Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Deemed to be University, Enathur, Kanchipuram, Tamil Nadu, India. His areas of interest include Mobile Application Development, Data Science and Data Analytics.