Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

IDENTIFYING PHISHING WEBSITES AND URLS A REAL-WORLD EXAMPLE UTILIZING DIFFERENT LOGIN URLS

[#1]MADUPU RAHUL,
[#2]P.SATHISH, ***Assistant Professor,***
[#3]Dr.V.BAPUJI, ***Associate Professor& HOD,***
Department of Master of Computer Applications,
VAAGESWARI COLLEGE OF ENGINEERING, KARIMNAGAR, TELANGANA

ABSTRACT: A web service is required for Internet communication software to function. The use of deceptive ways to steal personal information is on the rise. While convenient, it introduces numerous security flaws into the Internet's private infrastructure. One of several security concerns to web services is web phishing. Experienced users can detect phishing attacks, however novice users frequently prioritize security. Phishing is the practice of impersonating respectable website operators in order to steal sensitive user data. Phishing poses a significant risk to web security. Violent websites encourage internet crime and stifle the expansion of web services. As a result, there has been a significant effort to develop a comprehensive solution to ban the websites. Our idea is a literacy-focused strategy to categorizing webpages as benign, spam, or harmful. Our system only looks at URLs, not web page content. As a result, program stalling and drug users' web surfing dangers are reduced. Due to learning methodologies, our solution beats blacklisting services in generality and substance.

Keywords: Security; Web Services; URL; Vulnerabilities

1. INTRODUCTION

During the day, our primary concentration is on online work. Using a system and an internet connection in many ways simplifies both professional and personal life. This platform improves sales and operations in the commercial, medical, academic, information, financial, aeronautics, exploration, infrastructure, entertainment, and welfare sectors. Because of the advancement of mobile and wireless technology, drug users can now connect to a network and surf the internet 24 hours a day, seven days a week. Despite its ease, this system has revealed information security problems. As a result, online drug users must secure their computers. Data theft and other crimes can be committed by cybercriminals, hackers, and fair-limited users. The purpose is to access the system or its data in multiple methods, or to get specific data. Bushwhackers communicate with a variety of drug traffickers in order to gather knowledge and profit. According to Kaspersky, an attacker will cost between $108,000 and $1.4 billion by 2019, depending on the intensity of the attack. The billionaire spent $124 billion on global security products and services. Phishing is the most prevalent cybersecurity attack, and its perpetrators are cyber threats. Most victims are vulnerable to phishing attempts due to their lack of expertise about web operations, computer networks, and associated technology. It is easier to target drug addicts with fake websites and incentives to click on them than it is to breach the information security system. A malicious website imitates the original site's visual aesthetics and user experience by using copyrighted content such as the association's logos and other visual elements. Individuals and businesses have suffered considerable financial and reputational harm as a result of drug users unintentionally browsing phishing website URLs. The most dangerous cyberattack in this category was phishing. For this attack, cybercriminals use dispatch or other social media networks. Bushwhackers entice drug users

by pretending that the cash were transferred through a reliable online platform, such as an e-commerce site or comparable alternative, rather than traditional banking institutions. As a result, individuals attempt to obtain and retain crucial information from them. This is done by bushwhackers to discredit their targets' stories. As a result, it produces financial loss as well as long-term injury.

2. LITERATURE SURVEY

The population's maturity has been corrupted, causing people to accidentally give hackers their personal information. Numerous illegal websites have been formed to steal private information from marijuana users. Passcodes, savings accounts, and delivery details are examples. Since 2004, the group has documented an unusual number of hacking events in the latter part of 2016. In 2016, 1,609 phishing assaults were detected. The data from 2015 has increased by 65%. Fraud occurred each month in the fourth quarter of 2004. The phishing website was identified using Machine Learning. Grounded Malware Monitoring Systems use machine literacy to deliver functions. URLs, sphere names, online features, and website content are organized to form features.

Due to its nonlinear framework, the web security system can detect irregularities on numerous internet platforms, making it fashionable. Machine literacy properties are compared to URLs, key law sources, and third-party services. To detect such attacks, a way to test the trustworthiness of computers developed to recognize misleading conduct in internet communication by marijuana users is feasible.

This method can identify phishing websites and email textbook distribution schemes. S. Marchal et al. use authentic point garçon record data to identify fraudulent URLs. Disabling the operation or finding a dangerous website. Open source has many advantages, including intimate proximity, complete autonomy, extraordinary language flexibility, rapid adaptation, resistance to active phishing, and resistance to phishing method development. The bracket approach for spotting bogus sites was developed by Mustafa Aydin et al. This technique evaluates grounded point subset selection and rooted website URL properties. Phishing websites are detected using point birth and selection methods.

Fadi Thabtah et al. used a variety of machine learning methods to analyze authentic malware samples and manipulate experimental parameters. This comparative examination examines machine learning forecasting models' pros and cons and their phishing efficacy. Based on empirical evidence, covering approach models are better anti-phishing solutions. The Anti Phishing Simulator by Muhemmet Baykara et al. exposes phishing detection issues and provides suggestions for detecting bogus emails. The study recommends using "textbook of e-mail" only for complex word processing activities.

3. RELATED WORK

The linguistic traits come from the apparent URL inconsistencies between illegal and lawful websites. The lexical properties can capture quality during categorisation. Isolating URL hostname and path simplifies bag-of-words extraction. Phishing websites love long URLs with multiple situations, commemoratives in a field, and tokens.

Using well-known brand names as commemorative entities can make phishing and malware websites look legitimate. Sometimes phishing and malware websites replace the suspicious URL with the IP address, which is rare in benign situations. Phishing URLs sometimes use deceptive commemorative terms. To strengthen security, we store critical terms' double values in our feature set. Benign bones are preferable over aggressive bones. Thus, fashionability matters. Host-based features are employed because aggressive patches persist in hosting hubs or low-rated locales. The dataset URLs teach arbitrary timber and support vector machine supervised learning techniques.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

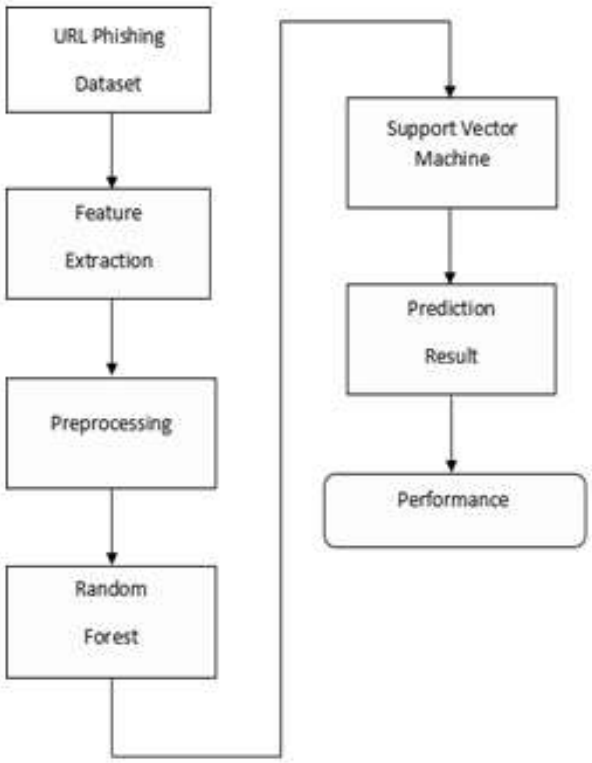

Fig 1. Architecture for proposed system

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.

4. METHODOLOGY

Modules:

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

DATA COLLECTION

The data utilized in the study consists of records. This phase entails the process of choosing a representative subset from the entirety of available data. The challenges associated with machine learning primarily arise from the data aspect, particularly the substantial volume of data required to ascertain the desired outcome. Labeled data is characterized by the presence of replies.

DATA PRE-PROCESSING

Formatting, cleansing, and sampling are methods of organizing selected data. In academia and research, three common data pre-processing methods are utilized.

Manipulation of the selected data will be challenging. It is desired to export relational database data to a flat file. If the data was in another format, it should be exported to a relational database or text.

Data cleansing is the process of removing and replacing missing data. In other circumstances, data may be incomplete, prohibiting us from addressing the issue effectively. All of these, most likely, must be abolished. Some attributes may include sensitive information that has to be cleared or deleted from the dataset.

Sampling: We might have more relevant data than we need. Additional information lengthens method execution times and raises computational and storage requirements. Prior to a comprehensive investigation, a smaller dataset sample may be used to accelerate concept discovery and development. This method simplifies data processing, allowing for faster concept formulation and refinement.

FEATURE EXTRACTION

Next is feature extraction, which involves attribute extension to create URL-based columns. Finally, a classifier algorithm trains our models. They benefit from the categorized dataset. The remaining categorized data from our study will validate the models. Machine learning is used to recognize pre-processed data. This study uses Random Forest classifier.

EVALUATION MODEL

The examination of models is critical to evolution. It is used to determine which model best describes the data and its predicted behavior. Model accuracy assessments for two independent approaches are required to prevent overfitting. The median efficiency of a classification model is measured. The outcome will be as expected. Graphics are used in classification. A percentage of testing dataset predictions that are correct. It is

simple to calculate by dividing total forecasts by accurately expected estimates. Accuracy is the difference between actual and expected production.

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

Where TP = True Positives TN = True Negatives

FN = False Negatives FP = False Positives

5. PROJECT DESCRIPTION

That data comprises various elements to evaluate when determining if a website URL is licit or phishing. Phishing website finding and bracketing criteria are as follows.

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

ADDRESS BAR BASED FEATURES

The use of IP addresses

If a URL contains an IP address instead of a domain name, such as 125.96.2.121, a person may suspect their sensitive data is being compromised.

The suspicious component is buried behind a lengthy URL.

Phishers can hide a suspicious URL in the URL bar by utilizing an extended URL. Long URLs are shortened by web tools to make them easier to manage. These services are often used to make URLs memorable and shareable. The URL is concise. Internet URL shortening shortens URLs while directing viewers to the proper page.

URLs use "@" symbol.

Web users often ignore everything before the @ symbol when navigating URLs because the real address follows.

Redirecting with a double slash

The URL path with double forward slashes takes users to another website. A hyphen (-) separates domain prefixes and suffixes.

Hyphens are rare in legitimate URLs. Fraudsters use hyphens to add prefixes or suffixes to domain names to make them appear valid.

Multiple subdomains are common in web hosting.

Consider visiting the URL below. http://www.kanchi.ac.in/students/ has student resources. Some legal specialties at the top of the country could be called sphere names.

HTTPS secures internet data flow.

HTTPS is essential for website validity, however it is not enough.

The term of domain registration is discussed.

Phishing websites are temporary, so safe practices must be followed often. Our sample has only one longest phony discipline.

A "favicon" is a web browser icon.

A website's favicon is usually a photo.

Using an unusual port

This functionality helps determine if a server service is overloading or unavailable.

An HTTPS token in a URL address Phishers may put a commemorative HTTPS element in a URL's circular component to confuse drug abusers.

ABNORMAL BASED FEATURES

URL Request Form

The URL test checks a website's external images, videos, and audio for their source.

An "anchor link" is a hyperlink that lets users easily jump to a certain page on a webpage.

Using tags, an anchor describes an element. The attribute is Request URL.

The user's text is scholarly without rewriting.

Our extensive study of website supply code showed that respectable websites use tags to deliver HTML page information. Use script tags to write user scripts. Link tags help you access many internet sites. All tags must be linked to a website page.

Server from Handler

The reported file needs processing, so SFHs with an empty string or a "clean" description are questionable.

Sending an email is a frequent way to communicate in modern life.

The application form lets clients enter confidential information that a server processes. Phishers may steal client data from their personal email accounts.

The URL is unusual.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

The WHOIS network supplies this data. The URL of a respectable website usually includes an identifier.

HTML AND JAVASCRIPT BASED FEATURES

Please translate the user's text to academic style without including any further information.

The frequency of page redirects appears to differentiate phishing websites from legitimate ones.

Disabling right-click Phishers restrict clients from accessing and purchasing website applications by limiting right-click capabilities with JavaScript. This function is used in conjunction with the "Using on Mouseover to Cover the Link" function.

Window with a Pop-Up Use

Malicious websites that request personal information through pop-up windows are uncommon.

IFrame Redirect IFrames allow websites to be integrated in the current page.

DOMAIN-BASED FEATURES

Age of the domain

The WHOIS network can extract this function. The majority of phishing websites are temporary. The eligible region must also be 6 months old, according to the data.

Phishing websites either cannot be verified using the WHOIS database or lack supporting proof. A website is Phishing if its DNS document is blank or inaccessible; otherwise, it is Valid.

Website traffic is the number of visitors to a website over time.

The function counts website visitors and pages seen to evaluate its performance.

PageRank is a market price metric with a value between 0 and 1. PageRank is a computational method that ranks websites by popularity or importance on the Internet.

Internet search using Google.

The function does not decide if Google should index a webpage. Upon Google registration, a domain is added to the list.

Number of links to a page

Although some links may be identical in size, the number of links to a website indicates its credibility.

6. RESULTS

Machine learning algorithms were imported by Scikit-learn. A training set is used for each classification, and a testing set is used to evaluate classifier performance. To assess performance, classifier accuracy was calculated.

Table 1. Dataset D1 text feature performance with different classifiers.

Classifier	Textual content features	Pre (%)	Recall (%)	F-Score (%)	AUC (%)	Acc (%)
LR	TF-IDF word level	85.68	88.25	86.95	85.38	85.62
	TF-IDF N-gram level	85.23	85.42	85.33	83.93	84.05
	TF-IDF character level	84.55	87.15	85.83	84.13	84.39
	Count vectors	86.84	79.12	82.80	82.45	82.16
	Word sequences vectors	55.87	83.27	66.87	52.61	55.23
XGBoost	TF-IDF word level	88.44	88.56	88.50	87.41	87.52
	TF-IDF N-gram level	87.77	86.51	87.13	86.10	86.14
	TF-IDF character level	**89.01**	**90.58**	**89.79**	**88.65**	**88.82**
	Word sequences vectors	82.66	85.87	84.23	82.24	82.55
	Count vectors	88.26	87.75	88.00	86.95	87.02
	Character sequences vectors	81.47	87.81	84.52	82.05	82.54
RF	TF-IDF word level	85.94	92.67	89.18	87.34	87.80
	TF-IDF N-gram level	86.77	89.57	88.14	86.68	86.93
	TF-IDF character level	85.44	92.81	88.97	87.02	87.51
	Count vectors	85.81	93.08	89.30	87.41	87.90
	Word sequences vectors	81.56	90.71	85.89	83.19	83.83
	Character sequences vectors	79.51	93.91	86.11	82.60	83.56
NB	TF-IDF word level	84.50	79.12	81.72	80.95	80.79
	TF-IDF N-gram level	82.45	71.16	76.39	76.59	76.13
	TF-IDF character level	76.45	81.89	79.08	75.98	76.49
	Count vectors	82.62	71.63	76.74	76.88	76.43
	Word sequences vectors	62.89	42.66	50.83	56.39	55.22
DNN	TF-IDF word level	87.08	91.20	89.09	87.57	87.88
	TF-IDF N-gram level	88.12	84.29	86.17	85.40	85.31
	TF-IDF character level	**88.32**	**91.62**	**89.94**	**88.62**	**88.40**
	Count vectors	87.49	89.46	88.47	87.14	87.34
	Word sequences vectors	54.26	100.0	70.35	50.0	54.26
	Character sequences vectors	76.41	91.43	83.25	78.97	80.03
LSTM	GloVe pre-trained word embedding	87.05	90.79	88.88	87.38	87.67
	Trained word embedding	88.14	89.20	88.66	87.48	87.62
CNN	Character embedding	82.06	89.34	85.54	83.08	83.61
	Trained word embedding	89.57	85.00	87.22	86.62	86.49

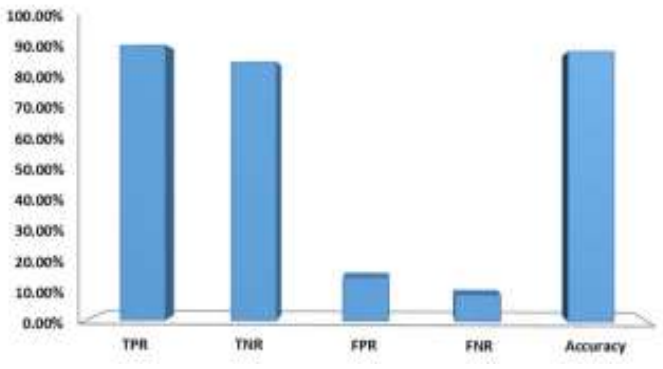

Figure 2. Textual content feature performance

Table 2. Performance of proposed hyperlink attributes on D1 with different classifiers.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 9, September : 2023

Classifier	Precision (%)	Recall (%)	F_Measure (%)	AUC (%)	Accuracy (%)
RF	77.59	86.10	81.63	82.57	82.27
Ensemble	77.39	86.23	81.57	82.50	82.18
LR	69.05	55.67	61.65	67.32	68.31
NB	68.31	31.60	43.21	59.62	62.01
XGBoost	75.55	84.77	79.90	80.82	80.49

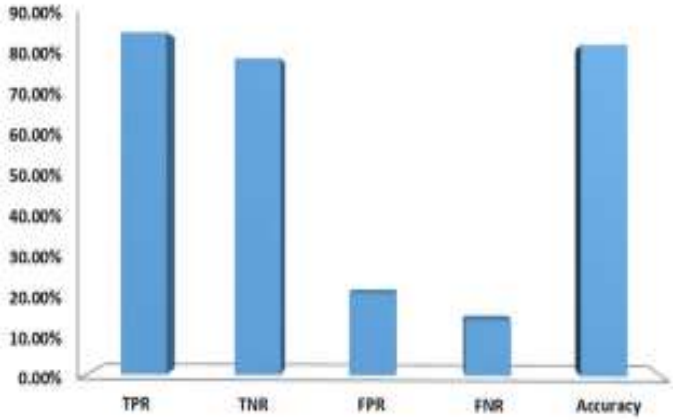

Figure 3. Performance of hyperlink based features

Table 3. Performance of several feature combinations on D1 with different classifiers

Classifier	Features	Pre (%)	Recall (%)	F-Score (%)	AUC (%)	ACC (%)
LR	F_{URL}	74.67	67.92	71.13	74.25	74.79
	F_{HTML}	83.50	81.98	82.74	84.16	84.35
	$F_{URL+HTML}$	77.71	68.74	72.95	76.06	76.68
NB	F_{URL}	81.41	22.09	34.76	58.92	62.06
	F_{HTML}	65.67	87.57	75.06	74.49	73.38
	$F_{HTM+URL}$	86.99	62.15	72.51	77.16	78.44
Ensemble	F_{URL}	98.42	92.05	95.13	95.40	95.69
	F_{HTML}	90.22	82.01	85.92	87.25	87.70
	$F_{URL+HTML}$	93.89	87.85	90.77	91.52	91.83
RF	F_{URL}	98.54	92.14	95.23	95.49	95.78
	F_{HTML}	90.77	81.98	86.16	87.48	87.95
	$F_{URL+HTML}$	93.81	86.79	90.16	90.98	91.34
XGBoost	F_{URL}	99.58	92.27	95.79	95.97	96.29
	F_{HTML}	88.21	87.68	87.94	88.90	89.01
	$F_{URL+HTML}$	**98.28**	**94.56**	**96.38**	**96.58**	**96.76**

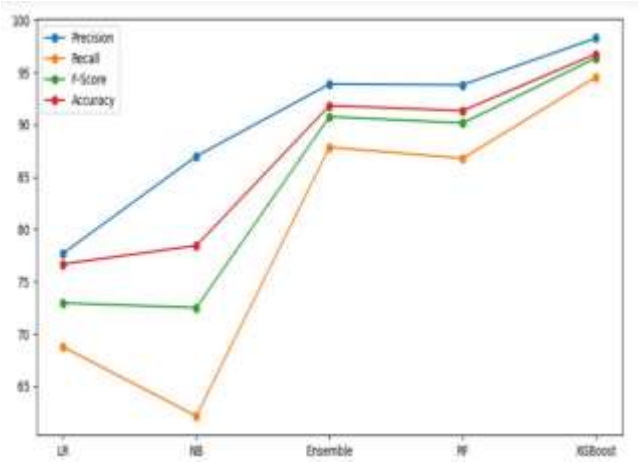

Figure 4. Different classifiers' combined feature test results

7. CONCLUSION

This study analyzes our powerful large-scale technology that automatically classifies phishing campaigns with a false positive rate of less than 0.1. Our bracket system easily reviews several implicit phishing runner responses in less time than a specialized review method. The classifier in our system automatically simplifies our blacklist, reducing the time phishing operators have to commit crimes before we take action. Our blacklist approach against phishing is effective due to a robust classifier and resilient infrastructure. Machine literacy can distinguish phishing and authentic URLs. We bought a delicacy meter.

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.