



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Detection of Phishing Websites using Machine Learning

Likhitha¹, Abhinaya A², D Poojitha Chowdary³, Devika Sm⁴

Associate Professor, Department of CSE, Nagarjuna College of Engineering and Technology, Bengaluru, India¹

U.G. Student, Department of CSE, Nagarjuna College of Engineering and Technology, Bengaluru, India^{2,3,4,5}

ABSTRACT: Phishing is one of the most common and dangerous kinds of cyberattacks. The goal of these attacks is to steal the information that individuals and companies need to complete transactions. Phishing websites include a variety of clues in their content and information derived from web browsers. A new type of network attack known as "phishing" involves creating a duplicate of an existing webpage in order to deceive users into divulging sensitive personal information, like passwords, to a website that purports to be the service provider (for instance, through carefully worded emails or instant messages). This project aims to use machine learning (ML) to identify features in the UCI Machine Learning Repository database, such as information regarding phishing websites.

I. INTRODUCTION

Data security in digital systems is becoming increasingly important due to the rapid growth of technology and the growing use of these systems. The basic objective of maintaining information technology security is ensuring that the right protections are put in place against dangers and hazards that users are likely to encounter when utilising these technologies. The term "phishing" first appeared in the 1990s. Early hackers, who mostly used phones to hack, occasionally swapped out the letter "f" for the letter "ph" to create new phrases in their network. Hackers use a tactic called "phishing," a play on the words "fishing," to pretend to be an email or instant messaging service provider and fool users into visiting dubious websites. Using this method, the victim's sensitive information, including passwords, user names, and national security IDs, can be accessed by an attacker. Therefore, in the future, these details might be utilised for targeted ads or identity theft schemes (such as transferring money out of the victims' bank accounts). The most common attack technique involves sending emails to potential victims that appear to be from internet companies, banks, or ISPs. The purpose of these emails is to convince you to visit the website by clicking on the provided URL, where you can change your password and account number. They will invent a multitude of excuses, such as saying that your credit card password has been input incorrectly multiple times or that they are providing updating services. When you click on these links, a phoney website will open.

II. LITERATURE SURVEY

Classification of Phishing Websites using Intelligent Rules:

Phishing is the practice of mimicking a trustworthy company's website design in order to get customers' personal information, such as passwords, social security numbers, and usernames. Phishing websites can be identified by their content in a number of ways and by security indicators that depend on the browser. Many strategies have been proposed to combat phishing. However, there isn't a single fix that can completely resolve this problem. According to data mining, this is one approach to anticipate phishing attacks.

In particular, the "induction of classification rules," since anti-phishing solutions closely coincide with classification data mining, which aims to forecast the type of website. In this work, we evaluate whether rule-based classification data mining approaches are appropriate for detecting phishing websites and we pinpoint the key features that distinguish authentic websites from fake ones.[1]

Evaluation of Phishing-Related Features Websites that make Use of Automation

Businesses that enable cross-border trade can gain a competitive advantage by catering to a worldwide customer base. Internet merchants encounter a variety of challenges, including insecure money orders. Phishing is the practice of impersonating a trustworthy company's website in order to get personal information such as social security numbers, usernames, and passwords. Phishing is regarded as an illegal online activity.

To distinguish trustworthy websites from fake ones, one can use prefixes and suffixes added to the domain and request URLs, extended URLs, IP addresses provided in URLs, and other features. In this work, we examine important features

that are automatically extracted from webpages using a novel technique, hence removing the need for a human expert to complete the extraction process. We also evaluate the features' importance in proving the website's validity. [2]

Multi-label classification rules for phishing

Associative categorization (AC) has few approaches because it is considered a difficult task to generate multi-label rules from single label data sets. Present AC methods only yield the greatest frequency class in the training data set that is linked to a rule, regardless of whether other classes have data representations related to the body of the rule. To address the aforementioned issue, we provide in this work the Enhanced Multi-label Classifiers based Associative Classification (eMCAC) algorithm, an AC method. Unlike other previous AC approaches, this method may infer rules associated with a set of classes from single label data. Furthermore, by employing a classifier building technique, eMCAC lowers the total number of extracted rules. The suggested algorithm's performance is tested on a real-world application data set pertaining to phishing websites, and the results show that eMCAC's accuracy is on par with other well-known AC and traditional data mining classification algorithms. Finally, our system's ability to derive new rules from phishing data sets and apply them to end users' decision-making is demonstrated by the experimental findings. [3]

A web-based system examination of phishing vulnerabilities

Nowadays, the most common technique employed by cybercriminals is phishing. Phishing is the illegal activity of sending phony emails and websites with false information in an effort to get personal, bank, and credit card details. Phishing attacks are always evolving in sophistication. Phishing has a detrimental effect since it raises the risk of identity theft and financial loss. Numerous institutions and associations are looking into this behavior and informing the public about the most recent strategies employed by the phishing sector.

Industry estimates indicate that there are more phishing attacks every year and that the anti-phishing technology available today are insufficient to identify phishing attempts. Furthermore, phishers always develop new, inventive techniques for their scams, making them harder to identify and stop. This essay offers a thorough description of the several phishing tactics, We create a fishbone diagram that explains the techniques and motives behind phishing after performing a root cause analysis of the motivation behind the phishing strategies. The goal of this research is to assist developers in building more effective anti-phishing systems. [4]

Web phishing page Detection using Anomalous Data

Recently, a lot of anti-phishing tactics have been put forth in the literature. In spite of all those precautions, phishing assaults nevertheless occur. The ability of phishing attackers to quickly and simply modify their strategies is one of the primary causes. In this study, we describe a novel technique that is not dependent on any particular phishing implementation. Our goal is to investigate the abnormalities found in online pages, namely the differences between the identity of a website and its HTML elements and HTTP operations.

Prior website expertise or user experience are not prerequisites. The cost of the attacker's attempt to get past our phishing detection will be high. [5]

III. SYSTEM DESIGN AND DEVELOPMENT

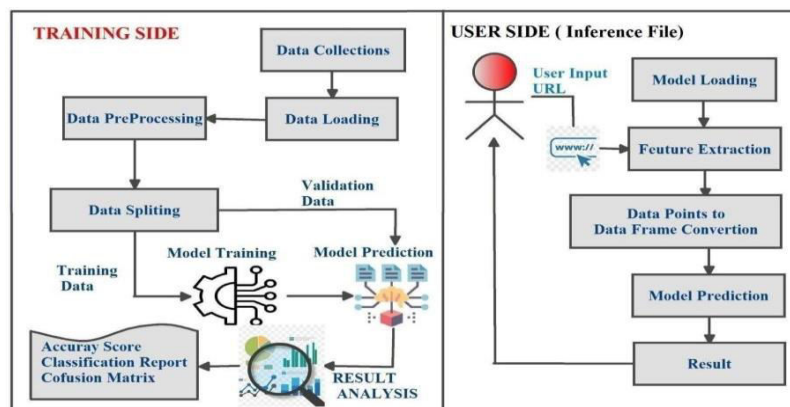
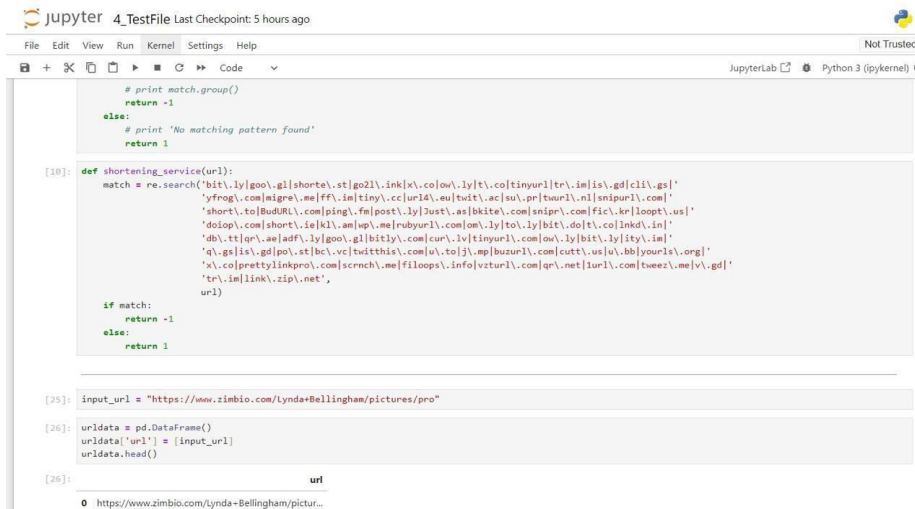


FIG 1: SYSTEM ARCHITECTURE

Framework The overall hypermedia structure of the WebApp is determined by its architecture design. Many factors, including the objectives set forth for a WebApp, the material to be shown, the anticipated user base, and the chosen navigational strategy, impact architecture design. How content items are structured for presentation and navigation is the main focus of content architecture.

WebApp architecture is concerned with how the application is configured to manage user interaction, carry out internal tasks, enable navigation, and show content. The WebApp architecture is defined with consideration for the development environment in which the application will be implemented.

IV. RESULTS



```

# print match.group()
return -1
else:
# print 'No matching pattern found'
return 1

[18]: def shortening_service(url):
match = re.search('bit\.ly|goo\.gl|shorte\.st|go2l\.ink|x\.co|ow\.ly|t\.co|tinyurl|tr\.im|is\.gd|cli\.gs|
'yfrog\.com|migre\.me|ff\|im|tiny\.cc|url4\.eu|twit\.ac|su\.pr|twurl\.nl|snipurl\.com|
'short\.to|BudURL\.com|ping\.fm|post\.ly|Just\.as|bkite\.com|snipr\.com|fic\.kr|loopt\.us|
'doioo\.com|short\.ie|kl\.am|up\.me|rubyurl\.com|ow\.ly|to\.ly|bit\.do|t\.co|lnk6\.in|
'do\|t|qr\.me|ad\.ly|goo\.gl|bitly\.com|cur\.lv|tinyurl\.com|ow\.ly|bit\.ly|tly\.in|
'q\.gs|is\.gd|po\.st|bc\.vc|twitthis\.com|u\.to|j\.mp|buzzurl\.com|cutt\.us|bb|yourls\.org|
'x\.co|prettylinkpro\.com|scrnch\.me|filoops\.info|vzturl\.com|qr\.net|1url\.com|tweez\.me|v\.gd|
'tr\.im|link\.zip\.net',
url)

if match:
return -1
else:
return 1

[25]: input_url = "https://www.zimbio.com/Lynda+Bellingham/pictures/pro"

[26]: urldata = pd.DataFrame()
urldata['url'] = [input_url]
urldata.head()

[26]:          url
0  https://www.zimbio.com/Lynda+Bellingham/pictur...
    
```

FIG 1: Giving URL as an Input



```

[[52 14 30 16 3 0 0 0 2 0 1 1 1 0 43 3 1 1]]

[29]: model_pred = model.predict_proba(urldata.values)
model_pred

[29]: array([[9.99520575e-01, 4.79424690e-04]])

[30]: model_pred.argmax()

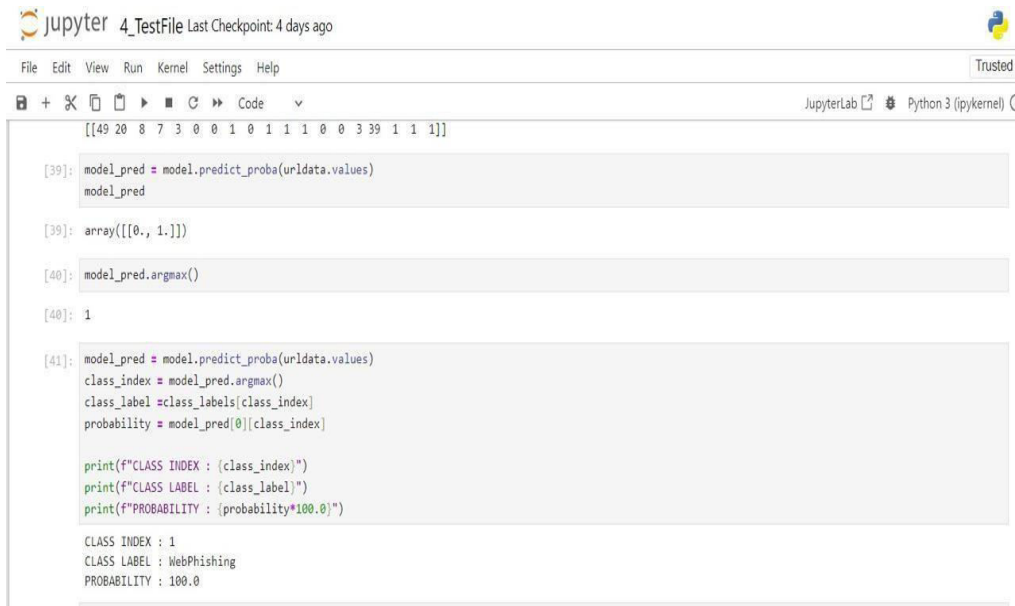
[30]: 0

[31]: model_pred = model.predict_proba(urldata.values)
class_index = model_pred.argmax()
class_label = class_labels[class_index]
probability = model_pred[0][class_index]

print(f"CLASS INDEX : {class_index}")
print(f"CLASS LABEL : {class_label}")
print(f"PROBABILITY : {probability*100.0}")

CLASS INDEX : 0
CLASS LABEL : Not WebPhishing
PROBABILITY : 99.95205753096285
    
```

FIG 2: NOT AN WEBPHISHING



```

[[49 20 8 7 3 0 0 1 0 1 1 1 0 0 3 39 1 1 1]]

[39]: model_pred = model.predict_proba(urldata.values)
      model_pred

[39]: array([[0., 1.]])

[40]: model_pred.argmax()

[40]: 1

[41]: model_pred = model.predict_proba(urldata.values)
      class_index = model_pred.argmax()
      class_label = class_labels[class_index]
      probability = model_pred[0][class_index]

      print(f"CLASS INDEX : {class_index}")
      print(f"CLASS LABEL : {class_label}")
      print(f"PROBABILITY : {probability*100.0}")

CLASS INDEX : 1
CLASS LABEL : WebPhishing
PROBABILITY : 100.0

```

FIG 3: WEBPHISHING

V. CONCLUSION

The goal of this study was to determine the most effective method for distinguishing phishing URLs from a sample of URLs that contained both benign and phishing URLs. Additionally, we have discussed host-based features, lexical analysis, statistical analysis, feature engineering, and dataset randomization for feature extraction. We also used a range of classifiers for the comparative analysis, and we found that the outcomes are almost the same for every classifier. Additionally, we saw that randomizing the dataset produced excellent optimization and a notable increase in the classifier's accuracy. We have chosen a straightforward method that makes use of basic regular expressions to extract the information from the URLs. Further features might be available for testing, which could increase the precision of the system even more. The dataset utilised for this work may have a somewhat outdated list of URLs; as a result, regular continuous training along with a new dataset would significantly enhance the model's accuracy and performance. The main problem with the content-based method of phishing URL detection is that phishing websites are hard to find, have brief lifespans, and make it difficult to train an ML classifier with their content-based characteristics. We refrained from utilising content-based features in our experiment due to these reasons. In the future, we want to incorporate a rule-based prediction that is derived from the content analysis of a URL. Therefore, a rule-based URL content analyzer and a classification-based lexical analyzer would be combined to provide a complete system for phishing URL detection.

VI. FUTURE SCOPE

Moving ahead, this study suggests various future directions. Enhancing accuracy by adapting models for dynamic phishing tactics and analyzing diverse data sources is important. Crucial areas include testing models across sectors, anomaly detection investigation, and addressing privacy concerns. Strengthening models against adversarial attacks enhances their resilience. Collaborative learning and user education are avenues to boost cyber security. Future research might focus on developing more sophisticated features that capture subtle nuances in phishing websites, such as analyzing the visual layout and design elements, dynamic content generation techniques, or behavioral patterns of users interacting with the site. In the current digital era, identifying phishing websites with machine learning is an essential responsibility. The prevalence of phishing attempts has increased, so it's critical to have high-accuracy algorithms to identify and stop them.

Finding phishing websites has shown to be a promising application of machine learning techniques. Using classification techniques, websites can be categorized as authentic or phishing, including supervised and unsupervised learning. Characteristics can be obtained from multiple sources, including as URLs, DNS, and data kept on websites such as Phishtank.

REFERENCES

1. L. McCluskey, F. Thabtah, and R. M. Mohammad. "Intelligent rule based phishing websites classification"(2014). IET Inf. Secure. 8(3):153–160.
2. R. M. Mohammad, F. Thabtah, and L. McCluskey. "Predicting phishing websites based on self-structuring neural network"(2014). Neural Comput. Appl.25(2):443–458.
3. N. Abdelhamid. "Multi-label rules for phishing classification"(2015).Appl. Comput. Informatics.11 (1): 29– 46.
4. W. D. Yu, S. Nargundkar, and N. Tiruthani. "A phishing vulnerability analysis of web based systems"(2008). IEEE Symp. Comput. Commun. (ISCC).326–331.
5. P. Ying and D. Xuhua. "Anomaly based web phishing page detection" (2006) in Proceedings - Annual Computer Security Applications Conference, ACSAC.381–390.
6. A. Sharan, "Rank fusion and semantic genetic notion based automatic query expansion model", Swarm and Evolutionary Computation, Vol-38, Elsevier, 2018.
7. Singh, J. "An Efficient Deep Neural Network Model for Music Classification", Int. J. Web Science, Vol. 3, No. 3, 2022.
8. Saurabh Kumar, S.K. Pathak, "A Comprehensive Study of XSS Attack and the Digital Forensic Models to Gather the Evidence". ECS Transactions, Volume 107, Number 1, 2022.
9. Anita S. Kini, A. Nanda Gopal Reddy, Manjit Kaur, S. Satheesh, Thomas
10. Martinetz, Hammam Alshazly, "Ensemble Deep Learning and Internet of ThingsBased Automated COVID-19 Diagnosis Framework", Contrast Media & Molecular Imaging, vol. 2022.
11. Ahmad Abunadi, Anazida Zainal ,Oluwatobi Akanb: Feature Extraction Process: A Phishing Detection Approach :In IEEE,2013



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details