International Journal of Integrated Science and Technology (IJIST)
Vol.2, No.1, 2024: 29-36

Malicious URL detection using Machine Learning

Prenalee Nanaware
All India Shri Shivaji Memorial Societys

Corresponding Author: Prenalee Nanaware pranaleenanaware@gmail.com

ARTICLEINFO

Keywords: Malicious URL Detection, Web Based Attacks, Feature Extraction, Malicious Web Pages

Received : 20, November
Revised : 21, Desember
Accepted: 29, January

ABSTRACT

One of the most prevalent and least protected security risks in existence today is fraudulent websites and URLs.We offer a method that both uses machine learning characteristics to identify phishing URLs and employs text processing techniques to evaluate text and identify incorrect remarks that are suggestive of phishing assaults.

DOI: https://doi.org/10.59890/ijist.v2i1.1289
https://journal.multitechpublisher.com/index.php/ijist

INTRODUCTION

Phishing attacks are when someone sends phony emails, texts, phone calls, or websites with the intention of tricking people into downloading malware, disclosing personal information (such as login credentials, bank account numbers, and credit card numbers), or taking other actions that put them or their companies at risk of being victims of cybercrime.

There are many developments in the field of technology for phishing detection. Heuristic techniques, feature-based machine learning techniques, and blacklist-based approaches are the three categories into which these are divided. While blacklists are well known for their excellent accuracy, they suffer from the time-lag effect, which makes it difficult for them to adapt to new phishing attacks. A heuristic or a machine learning system can identify recently established phishing websites across the nation. There are traits that are discovered to occur in phishing heuristic rules. While traits aren't always included in phishing assaults, they do exist in real-world instances.

When compared to machine learning algorithms, the heuristic approach is based on human experience rather than sophisticated data mining techniques. By choosing a set of discriminating features that could aid in differentiating between different sorts of websites, feature-based machine learning approaches attempt to discover a general mode to detect novel threats. Stated differently, characteristics are among the variables that can affect the accuracy of detection. Using position-based characteristics, we employ the Naïve Bayes algorithm in this strategy to minimize misjudgment.

The following characteristics are identified to determine if the white person is authentic or not: IP Address: An attacker can hide the true domain name of a website they are visiting by using IP addresses. URL Protocol: This section of the URL shows how detection algorithms can search the URL for an IP address to determine which network protocol should be used to retrieve the requested resource.

Since phishers are unable to utilize the precise URL they are targeting, they must create other pathways and domain names that resemble the original domain. This suggests that the URL may be a major factor in determining whether a website is phished. As a result, our method uses the URL to identify phishing websites. In particular, the y characteristics of the URL will be examined such that is able to utilize them to determine if the websites are phished.

In the last few years, a number of phishing solutions have been developed. These remedies include user education, government regulations prohibiting internet fraud, and technological safeguards. Reviewing previous studies helped us develop a more compressive research methodology for the current study and enhanced our fundamental understanding of the issue.

Machine Learning and Natural Language Processing for Phishing Attack Detection. This study presents a method that analyzes text using Natural

Language Processing techniques to identify remarks that are incorrect and suggestive of phishing attacks. This approach has been evaluated using a large benchmark collection of phishing emails in order to show how effective it is.

Detection of Malicious URLs using Machine Learning The purpose of this paper is to present a thorough analysis and a conceptual grasp of machine learning-based malicious URL detection methods. Traditionally, blacklists have been used to identify harmful online pages. However, machine learning techniques are employed to improve this because blacklists are not comprehensive and cannot identify freshly produced harmful URLs.

Three methods are used in the Heuristic-Based Approach to identify phishing: Heuristics, Machine Learning, and Blacklists

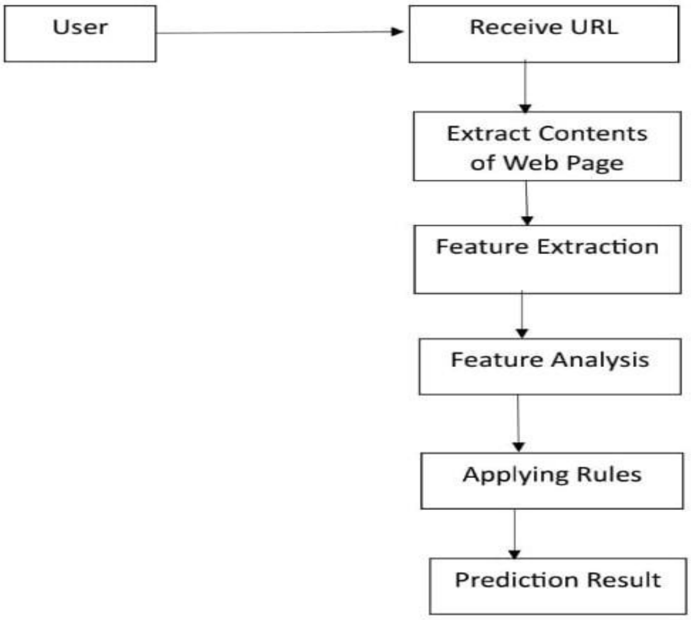
Phishing attack detection and prevention online. This research presents a novel end-host based anti-phishing algorithm called LinkGuard, which makes use of the common features of hyperlinks used in phishing assaults. The LinkGuard is capable of identifying both known and unidentified phishing attempts. It is lightweight and has real-time phishing attack detection and prevention capabilities.

Using Recently Registered Domains to Detect Phishing. In order to actively address the issue of phishing detection, this research proposes an Intelligent Phishing Detection (IPD) system. In particular, IPD uses the position-based feature-optimized Naive Bayes method to perform high precision detection after first automatically creating the detection dataset from the global huge domain name registration data.

METHODOLOGY

In our study, we have extensively addressed the intricate challenges in targeted display advertising through a carefully defined problem formulation. The fundamental obstacle we tackled is the cost and scarcity of training data from the target sampling distribution. To mitigate this, we introduced a two-stage transfer learning approach that harnesses models trained on surrogate domains and learning tasks and subsequently transfers this knowledge to the target task. Our empirical findings have underscored the remarkable value of different transfer stages in enhancing system performance. From these findings, several critical insights have emerged for the broader machine learning community.

These include the significance of deliberate data definition, the ability of transfer learning to combat cold-start problems, the importance of pragmatic constraints and data cost in decision-making, the efficacy of progressive dimensionality reduction, and the prevalence of transfer learning in diverse real-world applications. Overall, our study underscores the transformative potential of explicit transfer learning considerations in solving complex real-world challenges and guiding the development of automated systems.



SYSTEM DESIGN

The system model is depicted in following fig. 1:

1) Phase 1-Receiving URL of web page: The URL of web page is received from dataset or browser.

2) Phase 2-Contents of Web Page: In this phase, It will check whether the URL is valid or invalid.

3) Phase 3 - Feature Extraction: If URL is Invalid then Different features such as Length of URL, HTTP, HTTPS, symbols like ‘@’, ’-’ , ’//’. etc of URL are extracted to check whether URL is malicious or not.

4) Phase 4 - Feature Analysis: The features of URL are extracted and analyzed.

5) Phase 5-Applying Rules and Generation of Rules: The extracted features are analyzed and rules are generated to predicate results. The system architecture of proposed system as given below.

ALGORTITHM

```
Input:    URL
Output:  Predicted result whether it is valid or Invalid URL
Step 1: Start.
Step 2: Enter URL
Step 3: Extract the contents of web page in text format.
         href Count Script Count Symbol Analyze the result
Step 4: Extract images from the URL.
Step 5: Extract links from the given URL and write total count of the links.
Step 6: Analyze the URL features like Length, Symbols @,//,
         Https, Links, Scripts
Step 7. Apply the rules.
Step 8: Apply & Predict the result whether URL is Valid or Invalid
Step 9:  Stop.
```

EXPERIMENTAL RESULTS

We have selected some sites for training and testing Experimental procedure is divided into the following six phases:

A. Phase 1

Few websites a selected as shown in the Table

Tabel I: Some URL'S

SR.NO	URL
1	http://facebook.com
2	http://wallacefund.info
3	http://detdinsty.icu
4	http://fungi.myspecies.info
5	http://learn.knockoutjs.com

B.Phase 2

In this phase of content extractor, Count of three contents of webpage are extracted href, script and symbol-@ .

Tabel II: Three Features are Extracted

SR.NO	URL	Features		
		href	Script	Symbol
1	http://facebook.com	10	34	0
2	http://wallacefund.info	141	61	0
3	http://detdinsty.icu	0	2	0
4	http://fungi.myspecies.info	77	50	0
5	http://lesrn/knockoutjs.com	9	36	0

Table II shows the obtained result.

B.Phase 3

In this phase of Image extractor, the images from the given URL are extracted. Whatever images present in the site the sources of images are displayed in one side of image extractor. In this phase, on the other side also preview of the images of entered URL are displayed

C. Phase 4

In this phase of link extractor, the total number of links are displayed. The result is shown in the table III.

Tabel III: Total Number of links

SR.NO	URL	Totsl no of links
1	http://facebook.com	53
2	http://wallacefund.info	27
3	http://detdinsty.icu	21
4	http://fungi.myspecies.info	21
5	http://learn.knockout.com	71
6	http://learn.knockoutjs.com	14

D. Phase 5

In this phase of URL analysis, six features count are displayed.mThe result is shown in the table IV

Table IV: Six Features of URL are Analyzed

SR.NO	URL	Features				
		Length	@	//	HTTPS	https
1	http://facebook.com	25	0	1		
2	http://wallacefund.info	24	0	1		
3	http://detdinsty.icu	28	0	1		
4	Http://fungi.myspecies.info	28	0	1		
5	http://learn.knockoutjs.com	45	0	1		

E. Phase 6

In the final phase of prediction, from above count of total ten features the final result is calculated too predict whether the URL is valid or invalid with the help of positive and negative count as shown in the Table V.

Tabel V.Final Prediction

SR.NO	URL	Positive Count	Negative Count	Result
1	http://facebook.com	50	50	Valid
2	http://wallacefund.info	30	70	Invalid
3	http://detdinsty.icu	60	40	Valid
4	http://fungi.myspecies.info	30	70	Invalid
5	http://learn.knockoutjs.com	50	50	Valid

In this if the positive count is more than negative then URL is valid. If the positive count is less then negative The URL is Invalid and if the positive and negative count is similar then also it is consider as valid URL From this count the final result is displayed whether the entered URL is valid or invalid URL.

CONCLUSIONS AND RECOMMENDATIONS.

In conclusion, this paper offers valuable insights and practical lessons derived from a real-world, large-scale machine learning system for targeted display advertising. The system addresses the challenges of limited data availability by employing a two-stage transfer learning approach, leveraging different source sampling distributions and training labels before transferring the knowledge to the target task. Explicit consideration of the nuances in defining events (E), sampling distributions (P(E)), and labels (Y) can significantly enhance machine learning outcomes. Employing data from distributions and labels that differ from the target task can lead to performance improvements, highlighting the need for results adjustment to the target distribution

REFERENCES

J. Hong, “The state of phishing attacks,” Commun. ACM, vol. 55, no. 1,pp. 74–81, 2012.

D. Sahoo, C. Liu, and S. C. Hoi, “Malicious URL detection using machine learning: A survey,” 2017, arXiv:1701.07179.

E. Nowroozi, A. Dehghantanha, R. M. Parizi, and K.-K. R. Choo, “A survey of machine learning techniques in adversarial image forensics,” Comput. Security, vol. 100, Jan. 2021, Art. no. 102092.

“GitHub code.” Accessed: Apr. 7, 2022. [Online]. Available: https://github.com/ehsannowroozi/Sec_Classifying_URL_Detection

P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “PhishNet: Predictive blacklisting to detect phishing attacks,” in Proc. IEEE INFOCOM, 2010, pp. 1–5.

Y. He, Z. Zhong, S. Krasser, and Y. Tang, “Mining DNS for maliciousdomain registrations,” in Proc. 6th Int. Conf. Collaborative Comput. Netw., Appl. Worksharing (CollaborateCom), 2010, pp. 1–6.

Luong Nguyen, Minh Nguy and Ba Lam Tô "Detecting Phishing Web sites: A Heuristic URL-Based Approach The 2013 International Conference on Advanced Technologies for Communications (ATC'13).

Chen J. and Guo C. "Online Detection and Prevention of Phishing attacks", The First International Conference on Communications and Networking in China(oct 2006)

CRANOR, L, EGELMAN, S., HONG, J., AND ZHANG, Y. Phinding phish: An evaluation of anti- phishing toolbars. Tech. Rep. CMU-CyLab-06-018, Carnegie Mellon University CyLab, November 2006.

JAGATIC, TN, JOHNSON, N. A., JAKOBSSON, M., AND MENCZER, F. Social phishing. Commun ACM 50, 10 (2007), 94-100