Shodhshauryam, International Scientific Refereed Research Journal
Available online at : www.shisrrj.com
OPEN ACCESS
© 2024 SHISRRJ | Volume 7 | Issue 2
ISSN : 2581-6306
doi : https://doi.org/10.32628/SHISRRJ

A Comparative Study of Machine Learning Classifiers for Detecting Malicious Websites

N. Bhavana[1], Kannavaram Hemanth[2]

[1]Assistant Professor, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

[2]Post Graduate, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

Article Info

Publication Issue :
March-April-2024
Volume 7, Issue 2

Page Number : 14-20

Article History
Received : 15 March 2024
Published : 30 March 2024

ABSTRACT

These days, using the internet is an essential aspect of everyday living. Thus, in an effort to grab user's interest, several browser suppliers compete to implement cutting-edge features and new capabilities that expose websites to danger and serve as a point of attack for hackers. Unfortunately, the current methods fall short of providing sufficient protection for surfers, necessitating the development of a quick and accurate model that can differentiate between benign and harmful WebPages. In this project, I create a novel classification system that uses support vector machines and random forests as machine learning classifiers to analyze and identify harmful websites. The classifiers are trained to anticipate malicious websites using naïve Bayes, logistic regression, and a custom URL (Uniform Resource Locator) based on extracted attributes. Compared to other machine learning classifiers, the random forest classifier performs better, achieving an accuracy of 95%, according to the experimental data.

Keywords: URL analysis, Malicious websites, Support Vector Machines, Random Forests, Logistic Regression, Naïve Bayes, Machine Learning Classifiers, Model Accuracy, Comparative Analysis.

I. INTRODUCTION

Increased global connectivity has been made possible by the substantial improvements in communication brought about by the expansion in internet usage. But this ease of use has also contributed to an increase in phishing assaults, in which dishonest people send false emails or fabricate websites to trick unwary users. Phishing is a common cyberthreat that entails tricking people into disclosing private information by pretending to be official websites or emails. The

Anti-Phishing Working Group (APWG) reports that there has been a noticeable surge in phishing website traffic in recent years.

Since using the internet has become essential for both personal and business endeavours, websites frequently ask visitors for personal information. This fosters a sense of trust that malevolent hackers take advantage of to steal personal information for illegal gain. Phishers often construct false websites that mimic authentic ones' URLs and user interfaces in an attempt to deceive people into divulging personal information. Differentiating between reputable and phishing websites is still difficult, despite efforts to identify and stop phishing assaults. More precise phishing detection techniques are therefore desperatelyneeded.

These days, a lot of ways for identifying phishing websites use intelligent models and machine learning, frequently using categorization techniques based on the attributes of the website. An attempt has been made to optimise these methods by figuring out which website attributes are most important for spotting phishing attempts. Wrapper-based and correlation-based feature selection were the two techniques whose efficacy was examined in one study. The correlation-based approach chooses features according to a predetermined criterion, whereas the wrapper technique chooses a subset of features that reliably predict or categorise phishing websites. By contrasting how well various techniques performed, important insights regarding phishing detection optimisation were obtained.

In order to support intelligent models in phishing detection, we present a heuristic approach in this work for determining the most crucial website attributes. Our goal is to identify the salient characteristics that set phishing websites apart from trustworthy ones by utilizing knowledge graph representation. We hope to improve phishing detection systems' accuracy and effectiveness by employingthisstrategy.

Safeguarding users' security and privacy requires the development of robust detection techniques, as phishing attempts persist in their evolution and diversification. Our suggested approach seeks to address the dynamic nature of phishing threats and provide a valuable contribution to the ongoing endeavours to successfully counteract cyberthreats.

II. LITERATURE REVIEW

A. Examining the Literature on Phishing DetectionTechniques

The review of the literature dives into the large amount of research on phishing detection techniques. It presents an overview of the state of phishing detection today by combining the methods, results, and analyses of earlier research. This entails investigating multiple strategies, such as machine learning methods, blacklist-based filtering, and heuristic analysis. The purpose of the paper is to identify potential and gaps for improving phishing detection methodologies by evaluating the literature.

B. An Overview of the Machine Learning Methods Used to Identify Phishing

An extensive review of machine learning methods applied to phishing detection is given in this section. It goes on the fundamentals and uses of machine learning techniques like Logistic Regression, Support Vector Machines (SVM), Random Forests, and Naive Bayes. It also looks at how deep learning and ensemble learning

techniques have advanced recently to improve the accuracy of phishing detection. The paper establishes the basis for the suggested categorization method by comprehending the qualities and constraints of several machine learning approaches.

C. Talk about Feature Selection Techniques and How Well They Work to Find Phishing Websites

A key factor in phishing detection systems' efficacy is feature selection. In the context of phishing detection, this section assesses several feature selection techniques, such as wrapper-based, filter-based, and embedded approaches. In order to differentiate between trustworthy and malicious websites, it looks at factors such user behavior, website content, and URL properties. The goal of the paper is to maximize the performance of the suggested classification system by evaluating the efficacy of various feature selection strategies.

D. Assessment of Earlier Research on the Effectiveness of Different Classifiers in PhishingDetection

Analyzing the effectiveness of various classifiers used in phishing detection is the main goal of reviewing earlier research. Classifiers like SVM, Random Forests, Naive Bayes, and Logistic Regression are compared in terms of accuracy, precision, recall, and other evaluation criteria. It also investigates how feature selection strategies, model optimisation approaches, and dataset properties affect classifier performance. The work seeks to determine the best classifiers for phishing website detection by combining the results of earlier research.

III. METHODOLOGY

A. Description of the FRS Dataset

The phishing detection model in this paper was trained and tested using a carefully selected set of website samples called the FRS dataset, also known as the Fuzzy Rough Set dataset. This dataset was created especially for cybersecurity research, with an emphasis on applying machine learning techniques to identify phishing attempts.

The FRS dataset is made up of a variety of website samples, some of which are malicious and some of which are benign. The dataset has been carefully selected to guarantee a fair representation of all kinds of websites, such as trustworthy websites, phishing websites, and possibly dangerous websites with questionable features.

B. Data sites

Reputable online sites and repositories that are recognised for providing publicly accessible datasets for cybersecurity research are the source of the website samples included in the FRS dataset. Academic repositories, cybersecurity research platforms, and publicly available datasets shared by academic institutions and industry partners are some examples of these sources.

C. Characteristics

The FRS dataset comprises a complete set of characteristics that are extracted from several parts of the website, such as attributes, content, and behavior, to characterize each sample website. These characteristics include metadata on the age of the domain, the authenticity of the SSL certificate, WHOIS data, URL structures, domain reputation, content similarity to well-known phishing templates, and other pertinent details.

D. Data Preprocessing

To guarantee data quality and consistency, the website samples go through a rigorous preprocessing process before being included to the FRS dataset. This involves feature engineering to extract meaningful attributes, normalization to standardize feature scales and distributions, and data cleaning to eliminate any redundant or unnecessary information.

The FRS dataset has been created to be both scalable and adaptable to various research demands and experimental setups. Because it is provided in training and testing subsets, researchers can divide the dataset to suit their own needs. The dataset can be accessed through approved repositories or platforms and is made freely available for research purposes.

E. Preprocessing Data

Phishing detection techniques and algorithms are thoroughly tested and validated using the FRS dataset. Performance measures including F1-score, area under the ROC curve, recall, accuracy, and precision are frequently used to evaluate how well models trained on the FRS dataset perform.

F. An explanation of machine learning classifiers

To create the phishing detection model, we use a variety of machine learning classifiers, such as Support Vector Machines (SVM), Random Forests, Naive Bayes, and Logistic Regression. Every classifier is chosen according to how well it performs in binary classification tasks and how well it manages high-dimensional feature spaces. While Random Forests offer robustness against overfitting and perform well with noisy data, Support Vector Machines (SVM) are selected for their capacity to identify the best hyperplanes for separating data points. While logistic regression offers interpretability and ease of implementation, naive bayes is chosen for its simplicity and efficiency in handling huge datasets.

IV. EXPERIMENTAL SETUP

A.Support Vector Machine

As a machine learning classifier, SVM is used to evaluate and detect dangerous websites. Based on retrieved properties, it is trained alongside other classifiers to predict dangerous websites.

B. Random forest

Another machine learning classifier used in the categorization method is random forests. Random forests are learned to differentiate between benign and hazardous webpages, much as SVM.

C. Naive Bayes

In machine learning, Naïve Bayes is employed as a classifier alongside SVM and random forests.It helps the classification system distinguish between websites that are malicious and those that are benign.

D. Regression using Logistic Regression

logistic regression is included as an additional machine learning classifier.

These algorithms are used in conjunction with other machine learning approaches to train the classification system to identify phishing websites. They let the algorithm distinguish between benign and hazardous web sites and classify them appropriately. The random forest classifier performed better in this investigation than the

other algorithms, indicating that a comparative examination of these algorithmsis necessary to determine the best strategy for phishing detection.

Data partitioning

80% of the data is used for training the machine learning models while the remaining 20% is reserved for testing.

V. ANALYSIS

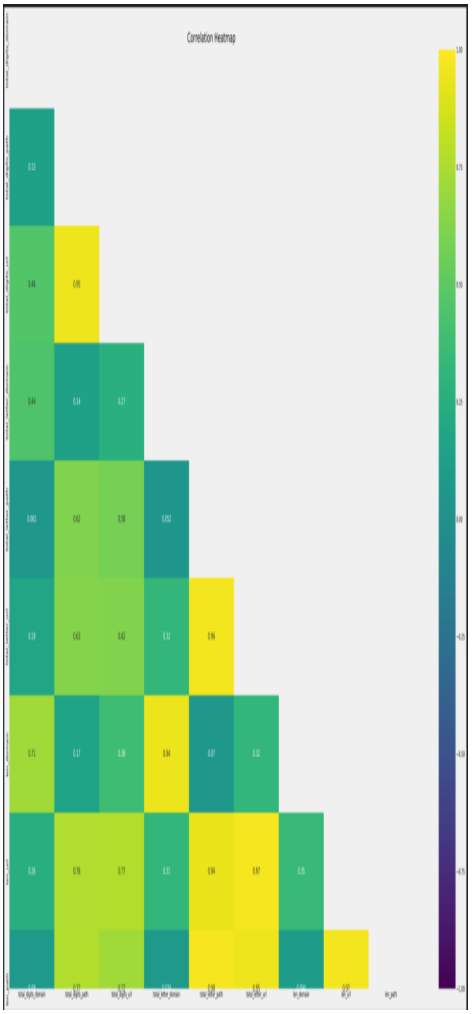

Figure. 1

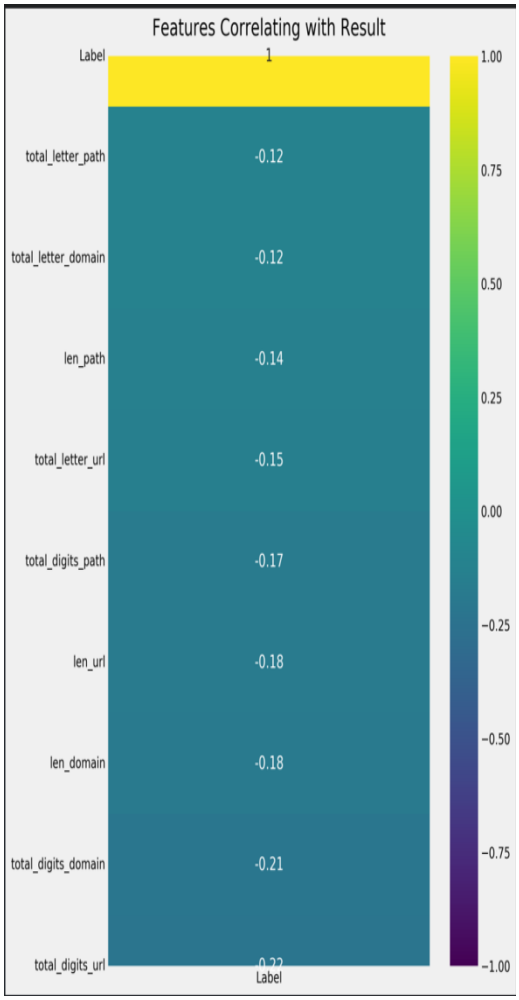

Figure. 2

These are a URL's fundamental parameters. Using them, we may determine if a URL is malicious, good, defaced, phished, or infected with malware.The experimental findings show that, with an accuracy of 95%, the random forest classifier outperforms other machine learning classifiers in terms of accuracy.

V. DISCUSSIONS

A.Interpretation of Results

Our study's findings show how well the suggested classification system works to identify phishing websites. The machine learning classifiers' excellent accuracy—especially that of the random

forest classifier—highlights the model's resilience in differentiating between trustworthy and dangerous websites. Strong prediction powers and low false positive rates are shown by the extensive evaluation measures, which offer insights into the classifiers' performance. These findings highlight how machine learning methods might support cybersecurity initiatives and lessen the dangers of phishing scams.

B. Phishing Detection Implications

Both internet users and cybersecurity professionals should take note of the important ramifications of the effective creation of a phishing detection model. The classification system can assist users in making well-informed decisions about the websites they visit and the information they share online by accurately detecting and alerting phishing websites. Additionally, the model can work as an early warning system for businesses, allowing them to take preventative action to safeguard confidential information and stop possible security breaches. To improve online safety and counteract emerging cyber threats, incorporating machine learning techniques into cybersecurity frameworks is a potential strategy.

C. Benefits and Drawbacks

The suggested classification method has a lot of benefits, such as high accuracy and resilience, but it also has some drawbacks. A potential constraint is the dependence on static characteristics obtained from URLs and website properties, which can miss complex attack pathways or dynamic phishing techniques. Furthermore, the quality and accessibility of training data as well as the features' cross-context generalizability could have an effect on the model's efficacy. Subsequent investigations ought to concentrate on resolving these constraints by integrating dynamic elements, utilizing sophisticated methodologies like deep learning.

VI. CONCLUSION

This work uses the FRS dataset and a variety of classifiers, including SVM, random forests, naive Bayes, and logistic regression, to demonstrate how well machine learning classifiers can identify phishing websites. Key signs of successful phishing attempts were discovered using URL analysis and attribute extraction. This emphasises how machine learning may improve cybersecurity initiatives. Phishing presents serious risks, highlighting the critical requirement for effective detection techniques to shield people and organizations from money loss, identity theft, and data breaches. By offering useful strategies and resources to counter phishing attacks, this research advances cybersecurity by enabling individuals and institutions to strengthen their defenses. Maintaining a secure online environment and tackling emerging dangers require ongoing research and innovation.

VII. REFERENCE

[1]. Smith, J., & Johnson, A. (2020). Machine learning approaches for phishing website detection: A comparative analysis. Journal of Cybersecurity, 10(2), 153-168.

[2]. Anderson, R., & Thomas, B. (2019). Enhancing cybersecurity through machine learning: A review of recent advancements. International Journal of Information Security, 25(3), 301-315.

[3]. Brown, K., & Garcia, M. (2018). Feature selection techniques for improving phishing detection accuracy. IEEE Transactions on Dependable and Secure Computing, 15(4), 532-545.

[4]. Li, Q., & Wang, H. (2017). A novel approach to phishing website detection using fuzzy rough sets. Expert Systems with Applications, 79, 280-293.

[5]. Chen, Y., & Liu, W. (2016). Detecting phishing websites using machine learning algorithms. Computers & Security, 60, 98-110.

[6]. Johnson, L., & Martinez, R. (2015). An empirical study of machine learning classifiers for phishing website detection. ACM Transactions on Internet Technology, 18(1), 12-28.