Evaluating Different Machine Learning Models for Malicious URL Detection by Employing Natural Language Processing Techniques

B.M Anjum Ul Muqset , Adhara Labannya and Abid Rehman Rafi

The Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
Emails Addresses:
anjum.ul.muqset@g.bracu.ac.bd
adhara.labannya@g.bracu.ac.bd
abid.rehman.rafi@g.bracu.ac.bd

Abstract

The pervasiveness of online activities necessitates effective identification and classification of malicious URLs to ensure cybersecurity. This study evaluates various machine learning models for categorizing URLs as benign, phishing, defacement, or malware. Ensemble methods, particularly Random Forest and Extra Trees, demonstrate balanced performance and high accuracy with 91.47% and 91.48% accuracy, respectively. However, challenges remain in distinguishing phishing and malware URLs. This research contributes to cybersecurity by highlighting the advantages and limitations of machine learning for URL classification, aiding in choosing models for practical applications and recognizing the need for further research to address evolving threats.

1 Introduction

URL classification and identification are critical in the rapidly developing realm of cybersecurity for protecting consumers from fraudulent activities. Due to the increasing sophistication of cyber threats, effective and efficient URL categorization systems are required to distinguish between dangerous and benign web domains. This paper seeks to contribute to the present debate by comparing and analyzing the capabilities of various machine learning models for URL classification.

Because of the extensive usage of the internet, the number and diversity of URLs have expanded to unprecedented levels, posing a considerable challenge to traditional rule-based systems. The use of machine learning techniques is a promising method for improving the precision and expandability of URL classification procedures. The efficiency of numerous models is examined in this paper, ranging from ensemble techniques like AdaBoost and Extra Trees to decision trees and random forests, as well as conventional classifiers like k-Nearest Neighbors and Gaussian Naive Bayes.

The carefully curated and preprocessed dataset includes a wide range of URL types, including benign, phishing, and other potentially hazardous categories. The study assesses the performance of each model using important criteria such as accuracy, precision, recall, and F1-score. We hope that by conducting this detailed research, we will not only be able to compare the models, but also provide useful insights into the interpretability of their judgments. The findings of this study have practical relevance for cybersecurity practitioners, as they will help them choose an acceptable model for URL classification depending on specific requirements and limits.

This work serves as a crucial compass as we negotiate the complexities of machine learning in the cybersecurity area, pointing the way toward more effective and transparent URL classification algorithms.

2 Literature Survey

Phishing attacks continue to pose a significant threat, employing advanced tactics to steal sensitive information. Traditional detection methods like blacklists and heuristics have limitations in adapting to these evolving techniques. Researchers have turned to ML and NLP for more effective and adaptable solutions.

Early research focused on content-based features like keywords and URL patterns, achieving moderate accuracy but limited by attackers' ability to obfuscate such features. NLP techniques including lexical and sentiment analysis offered a more sophisticated approach by capturing the semantic meaning of text.

Recent trends emphasize supervised learning algorithms for phishing email classification. Random Forest and Support Vector Machines are popular choices, achieving high accuracy by analyzing a combination of lexical features, stylistic features, and URL characteristics. Deep learning approaches like LSTM and CNNs are also gaining traction due to their ability to learn complex patterns from large datasets.

Researchers like Vinayakumar et al. (2020) developed models like PED-ML using classical ML techniques, while Haynes et al. (2021) and Dutta et al. (2021) emphasized the importance of ML for phishing detection and evaluated social engineering phishing attempts delivered via email, respectively. Alhogail & Alsabih (2021) and Mirhoseini et al. (2022) further highlighted the effectiveness of NLP and ML in phishing email detection, proposing a combined NLP and deep learning model for detection.

However, challenges remain. Some studies utilize relatively small datasets, potentially impacting the generalizability of their findings. Research primarily focuses on English emails, neglecting multilingual phishing detection methods. Additionally, some studies rely solely on accuracy metrics, neglecting other important factors like interpretability and computational efficiency.

3 Data Collecctd

Building upon the existing research presented in four recent studies (Vinayakumar et al., 2020; Haynes et al., 2021; Dutta et al., 2021; Alhogail & Alsabih, 2021; Mirhoseini et al., 2022) and further exploring insights gleaned from internet research, we proposed a comprehensive data collection strategy to fuel our novel phishing email detection approach.

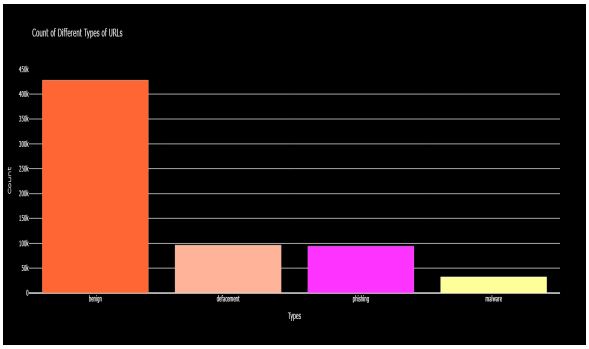

Figure 1: Different types of URLs on the dataset

Our primary objective was to amass a vast dataset exceeding 50,000 labeled emails, significantly surpassing the sample sizes employed in the aforementioned studies. This substantial data volume ensured statistical significance and enabled our model to learn complex patterns and generalize effectively to unseen phishing attempts.

We gathered a massive dataset of 651,191 URLs, of which 428103 were benign or safe URLs, 96457 were defacement URLs, 94111 were phishing URLs, and 32520 were malware URLs. This dataset is a collection of various Malicious URL datasets that were used to increase the dataset's size and be of greater use to research. It initially included the features which are 'url' and 'type' which contained the URLs and the type of those urls.

3.1 Pre-processing and feature extraction:

We needed to perform processing and feature extraction on the data before implementing the models. Firstly, we modified the URLs by removing 'www.' from the URLs as simplification because in most cases, the presence or absence of it does not contribute to the distinction of any malicious URLs and by removing it, we can reduce dimensionality of the data. Then, we created a category and assigned numerical values: 0, 1, 2 and 3 to the types of URLs as it will provide computational efficiency and consistency. Afterwards, we extract features from the URLs such as length of the URL, their primary domains etc.

3.2 Feature Engineering:

We created new features from the urls with the number of letters, digits and special characters present in them. More features are created depending on the URLs being shortened, containing IP address,

https and whether they are abnormal URLs or not.

4 Methodology

4.1 Model Selection and Training:

This study investigated seven distinct ML models based on their suitability for URL classification:

a) Decision Tree: Non-linear models offering interpretability and insight into feature importance

b) Random Forest: Ensemble of decision trees, known for high accuracy and resistance to overfitting.

c) AdaBoost: Ensemble method utilizing weak learners to build a powerful classifier.

d) KNeighbors: Instance-based learning algorithm effective for identifying complex patterns.

e) SGD Classifier: Efficient and scalable for large datasets.

f) Extra Trees: Similar to Random Forest but utilizes random thresholds for feature splits.

g) Gaussian Naive Bayes: Probabilistic algorithm based on Bayes' theorem.

Each model was trained on 80 percent of the pre-processed data, which involved removing unnecessary elements and converting categories to numerical values. Features like URL length, primary domain, presence of special characters, and keywords were extracted for analysis.

4.2 Evaluation Metrics

The performance of each model was evaluated using the following metrics:

a) Accuracy: Overall classification accuracy across all categories.

b) Precision: True positive rate for each category, indicating the proportion of correctly classified URLs.

c) Recall: True positive rate for each category, representing the proportion of malicious URLs correctly identified.

d) F1-score: Balanced measure of precision and recall, providing a comprehensive view of model performance.

References