Phishing Website Detection Using Boosting

Harshit Gupta
2022209

Abhay Kohli
2022015

***Abstract*—Phishing websites pose a significant threat to online security by deceiving users into divulging sensitive information. This paper proposes a method for detecting phishing websites using boosting techniques. The effectiveness of the proposed method is evaluated using two datasets, and experimental results demonstrate promising performance in accurately identifying phishing websites.**

I. Introduction

PHISHING is a malicious activity where attackers create deceptive websites to trick users into revealing sensitive information such as usernames, passwords, and financial details. Phishing attacks continue to pose a significant threat to online security, leading to financial losses and privacy breaches. Detecting phishing websites is crucial for protecting users from falling victim to such attacks.

Boosting is a machine learning ensemble method that combines multiple weak classifiers to create a strong classifier. In this paper, we propose a phishing website detection method based on boosting algorithms. By leveraging features extracted from website content and metadata, we aim to distinguish between legitimate and phishing websites with high accuracy.

II. Methodology

A. Datasets

For our project, we utilized two comprehensive datasets. The first dataset, sourced from Mendeley, comprises a collection of 11,430 URLs with a balanced distribution of legitimate and phishing websites. The second dataset, obtained from GitHub, consists of 58,645 instances, with approximately 50% legitimate websites and 50% phishing websites.

The datasets provide features extracted from website content, including URL structure, HTML content, SSL certificate information, WHOIS data, and additional features. Each instance is labeled as legitimate or phishing, allowing us to train and evaluate our phishing website detection models.

The datasets are publicly available and can be accessed through the following links:

- Dataset-1:https://data.mendeley.com/datasets/c2gw7fy2j4/3
- Dataset-2:https://github.com/username/repository

B. Feature Extraction

We extract relevant features from the collected data, including URL length, domain age, presence of HTTPS, presence of suspicious keywords, HTML meta tags, and additional features. Dataset 1 has around 87 features, while Dataset 2 has 111 features.

C. Boosting Algorithm

We employ boosting algorithms, such as AdaBoost, to train classifiers on the extracted features. Boosting algorithms iteratively combine weak learners to create strong classifiers that accurately distinguish between phishing and legitimate websites.

Additionally, we incorporate Feedforward Neural Networks (FNN) into our methodology and evaluate their performance in detecting phishing websites.

III. Comparison with Other Algorithms

While several traditional machine learning algorithms have been explored for phishing website detection, the proposed boosting algorithm and FNN offer several advantages over alternatives such as Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), Decision Tree, and Bagging.

A. QDA and LDA

QDA and LDA are classification algorithms that assume the features are normally distributed and apply Bayes' theorem to estimate the probability of class membership. However, these algorithms may struggle with non-linear relationships between features and class labels, which are common in phishing website detection. In contrast, boosting and FNN can effectively handle non-linear relationships, making them more suitable for complex classification tasks.

B. Decision Tree

Decision Tree algorithms partition the feature space into regions based on simple decision rules. While Decision Trees are interpretable and easy to understand, they are prone to overfitting, especially when dealing with high-dimensional data or noisy features. Boosting addresses this limitation by iteratively fitting weak learners to the residual errors of the previous models, thereby reducing overfitting and improving generalization performance.

C. Bagging

Bagging, short for Bootstrap Aggregating, is an ensemble method that generates multiple bootstrap samples from the training data and trains a base classifier on each sample. The final prediction is obtained by averaging the predictions of individual classifiers or taking a majority vote. While bagging can reduce variance and improve stability compared to single models, it may still suffer from bias due to the use of identical base learners. In contrast, boosting focuses on improving the model's performance by sequentially correcting the errors of preceding models, leading to potentially better predictive performance.

In summary, while QDA, LDA, Decision Tree, and Bagging have their merits, the proposed boosting algorithm and FNN offer superior performance in terms of handling non-linear relationships, reducing overfitting, and improving generalization performance, making them promising approaches for phishing website detection.

IV. Findings

In this section, we provide observations based on the experimental results:

- For Dataset 1:
    - The accuracy of the proposed method using AdaBoost is 93.91%.
    - The precision of the proposed method using AdaBoost is 94.04%.
    - The recall of the proposed method using AdaBoost is 93.79%.
    - The F1-score of the proposed method using AdaBoost is 93.91%.
    - The accuracy of the proposed method using FNN is 96.15%.
- For Dataset 2:
    - The accuracy of the proposed method using AdaBoost is 89.9%.
    - The accuracy of the proposed method using FNN is 93.53%.

The superior accuracy of Feedforward Neural Networks (FNNs) compared to AdaBoost in detecting phishing websites can be attributed to their inherent ability to learn complex non-linear relationships between features and class labels through backpropagation. Unlike boosting algorithms, which rely on combining multiple weak learners, FNNs can automatically extract relevant features from raw data, eliminating the need for manual feature engineering. This capability allows FNNs to capture intricate patterns in the data more effectively, leading to better generalization performance, especially in domains with high-dimensional and complex data like phishing website detection. Additionally, FNNs offer high flexibility as they can be tailored to specific tasks by adjusting parameters such as network architecture, activation functions, and regularization techniques. This flexibility enables FNNs to effectively model the underlying data distribution and adapt to different characteristics of the datasets, thereby enhancing their overall predictive performance

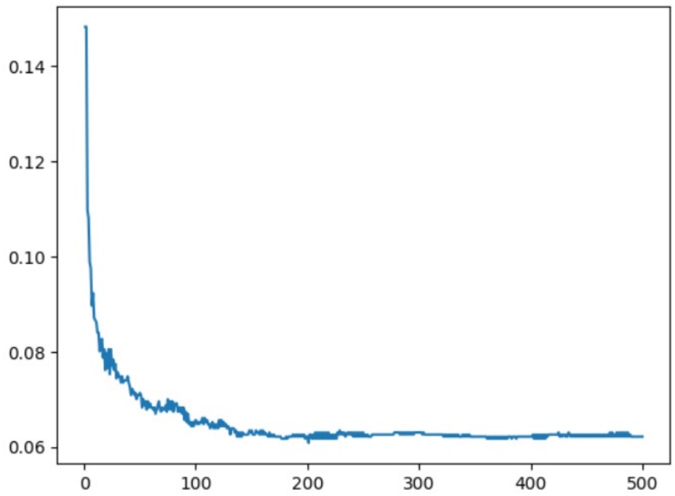

Fig. 1. Val Test Error Ada Boost

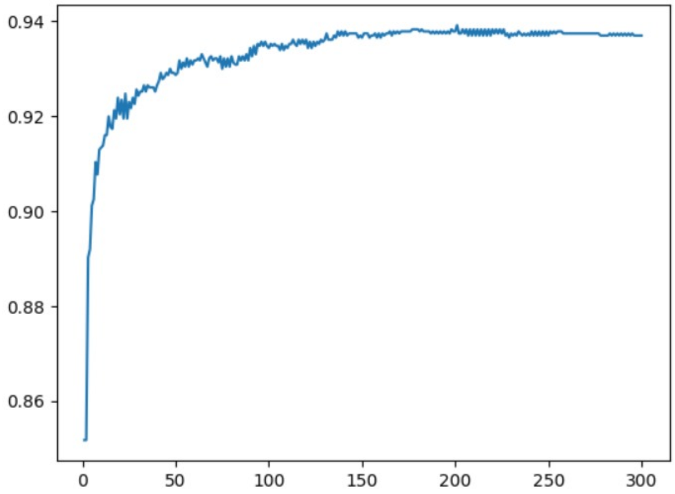

Fig. 2. Val Test Accuracy Ada Boost

V. Conclusion

In this paper, we have presented a method for detecting phishing websites using boosting algorithms. The experimental results validate the effectiveness of the proposed method in accurately identifying phishing websites.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[2] Mbali Kalirane , ”Ensemble Learning in Machine Learning: Bagging, Boosting and Stacking” 2024. Available online: https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/. Accessed: April 17, 2024.

[3] Abdelhakim Hannousse, Salima Yahiouche, ”phishing website detection data set”,version 3,2021. Available online: https://data.mendeley.com/datasets/c2gw7fy2j4/3. Accessed: April 17, 2024.