Mid-Term Project Report

Project Title: Phishing website detection using NLP models
Group ID: 25HR07

Table 1: Group Members

Name	Roll Number	Role / Responsibilities
Bharath Nayak	2301CS11	Meeting Scheduling, Documentation, Data gathering, Model Training Supervisor
Srikant Sahoo	2302CS07	Literature Survey, Web Scrapper, Backtesting, Model Evaluation, Hyperparameter Tuning
Chirag Ashish Agrawal	2301CS92	Model Implementation, Literature Survey, Deployment, EDA

1 Introduction

1.1 Overview of the Problem Domain

Phishing is a cybersecurity threat where attackers trick users into revealing sensitive information by imitating trusted websites. Traditional detection methods struggle with evolving phishing techniques. This project uses NLP to analyze website text for real-time phishing detection, combining cybersecurity and machine learning approaches.

1.2 Importance and Relevance

Online data is critical for identity, finance, and social life. Phishing attacks steal sensitive information, causing financial loss and privacy issues. Despite improved security tools, phishing continues to grow, making detection an important and timely cybersecurity challenge.

1.3 Scope of Our Project

The project detects phishing websites using NLP and traditional ML techniques. We analyze multiple aspects, URL, page structure (DOM), and text content, combining lexical, structural, and semantic features to identify phishing patterns. The goal is a model that effectively detects phishing websites in real time.

2 Related Work

Aleroud and Zhou [1] presented a detailed taxonomy of phishing detection methods, classifying them according to attack techniques, target platforms, and defensive mechanisms. Their survey emphasized the evolution from simple heuristic filters to more adaptive machine learning models and highlighted the need for hybrid systems that combine behavioral, content-based, and blacklist approaches to improve detection accuracy.

Kalla and Kuraku [2] concentrated on website-based phishing detection using URL features. They demonstrated that classical machine learning models such as decision trees, random forests, and support vector machines can effectively distinguish between phishing and legitimate websites when trained on lexical and statistical characteristics of URLs, including length, presence of special symbols, and domain-related attributes.

Rao et al. [3] introduced the use of word embeddings to represent textual content from webpages, capturing semantic relationships between words to enhance phishing detection. Their work showed that integrating these embeddings into models like logistic regression and neural networks improved the system's ability to identify subtle linguistic cues associated with phishing intent.

Aljofey et al. [4] proposed integrating HTML structural information with URL-based features to build more comprehensive machine learning models for phishing detection. By combining the hierarchical patterns of webpage layout with conventional lexical indicators, their method achieved greater robustness and classification accuracy across diverse phishing website datasets.

3 Methodology

3.1 Dataset Description

This study employs the Mendeley Web Page Phishing Detection dataset [5], which comprises 11,430 URLs with 87 extracted features, designed as a benchmark for machine learning-based phishing detection systems. The dataset is balanced, containing exactly 50% phishing and 50% legitimate URLs, and features are drawn from three categories: 56 based on URL structure and syntax, 24 from the content of corresponding pages, and 7 obtained by querying external services. The dataset is divided into part A and part B. Part A contains the URL and DOM of all websites. Part B contains the 87 extracted features.

3.2 Block Diagram

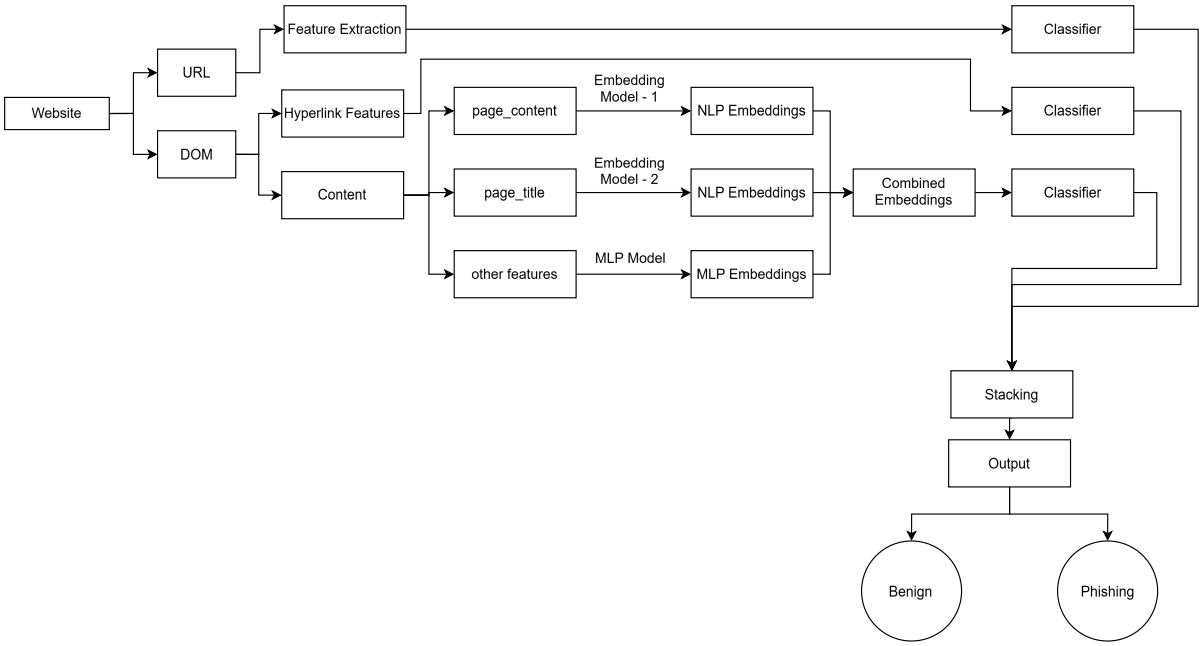

Figure 1: Overall project architecture.

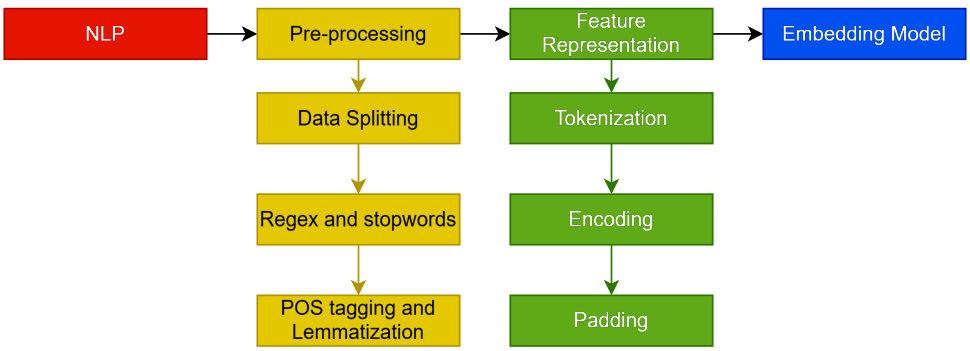

Figure 2: NLP workflow

3.3 Proposed Model

The proposed model adopts a three-branch ensemble architecture, where each branch is specialized to capture distinct aspects of a website for phishing detection. The URL branch focuses on extracting lexical and structural features from the website address, such as length, use of special characters, and subdomain patterns, which are indicative of obfuscation or malicious intent. The DOM/hyperlink branch analyzes the HTML structure and hyperlink relationships within the page, computing features like the ratio of internal to external links and the presence of suspicious scripts or redirects. The content-NLP branch leverages Transformer-based embeddings to encode semantic information, word embedding model for page_title and contextual embedding model for page_content.

Each branch independently produces a probabilistic prediction, which is then aggregated by a stacking meta-learner, typically a logistic regression model trained on out-of-fold base predictions. This meta-learner synthesizes the complementary strengths of the individual branches, yielding a final benign or phishing classification. The ensemble design enhances robustness to missing or adversarially manipulated features and consistently outperforms single-view models by integrating lexical, structural, and semantic evidence for comprehensive phishing detection.

3.4 Mathematical Model

Log Loss for classifiers with regularization:

$$L(\Theta) = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big] + \lambda \sum_{l} \|\theta_l\|_2^2 \tag{1}$$

Gradient of the loss:

$$\nabla_\Theta L = -\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)x_i + 2\lambda\Theta \tag{2}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{F1 - Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

$$\text{Mathews Correlation Coefficient} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{7}$$

Where:

Θ**:** Training parameters (weights) of the model

N**:** Number of training samples

y_i**:** True label of the i-th sample

$\hat{y}_i$**:** Predicted probability of the i-th sample

λ**:** Regularization coefficient

θ_l**:** Parameters of the l-th layer

x_i**:** Feature vector of the i-th sample

3.5 Algorithm

Algorithm 1 Stacked Ensemble for Phishing Website Detection

Require: Website dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ ($x^{(i)}$ has URL, DOM, and page text)
Ensure: Final meta-model M and base models B_u, B_d, B_t

Preprocessing:
for each sample x **do**
    $u \leftarrow$ url_features(x_{url})
    $d \leftarrow$ dom_hyperlink_features(x_{dom})
    $e_t \leftarrow E_1$(page_title) ▷ Transformer embedding
    $e_b \leftarrow E_2$(page_content) ▷ Transformer embedding
    $c \leftarrow$ other_content_features
    $t \leftarrow$ concat(e_t, e_b, c)
end for
Train base learners with K-fold OOF:
for $k = 1$ **to** K **do**
    Train B_u on $u[train_k]$; predict $oof\_p_u[val_k]$
    Train B_d on $d[train_k]$; predict $oof\_p_d[val_k]$
    Train B_t on $t[train_k]$; predict $oof\_p_t[val_k]$
end for
Fit meta-learner:
$S =$ concat($oof\_p_u, oof\_p_d, oof\_p_t$) ▷ Meta-features
Train M (e.g., Logistic Regression) on $S \rightarrow y$
Inference for new website x^*:
$u^*, d^*, t^* \leftarrow$ featureize(x^*)
$p^* = [B_u(u^*), B_d(d^*), B_t(t^*)]$
$\hat{y} = M(p^*)$
return ($\hat{y} \geq \tau$)

References

[1] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers amp; Security*, vol. 68, p. 160–196, Jul. 2017. [Online]. Available: http://dx.doi.org/10.1016/j.cose.2017.04.006

[2] D. Kalla and S. Kuraku, "Phishing website url's detection using nlp and machine learning techniques," *Journal on Artificial Intelligence*, vol. 5, no. 0, p. 145–162, 2023. [Online]. Available: http://dx.doi.org/10.32604/jai.2023.043366

[3] R. S. Rao, A. Umarekar, and A. R. Pais, "Application of word embedding and machine learning in detecting phishing websites," *Telecommunication Systems*, vol. 79, no. 1, p. 33–45, Nov. 2021. [Online]. Available: http://dx.doi.org/10.1007/s11235-021-00850-6

[4] A. Aljofey, Q. Jiang, A. Rasool, H. Chen, W. Liu, Q. Qu, and Y. Wang, "An effective detection approach for phishing websites using url and html features," *Scientific Reports*, vol. 12, no. 1, May 2022. [Online]. Available: http://dx.doi.org/10.1038/s41598-022-10841-5

[5] A. Hannousse, "Web page phishing detection," 2020. [Online]. Available: https://data.mendeley.com/datasets/c2gw7fy2j4/2