PARSHWANATH CHARITABLE TRUST'S

A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering

Data Science

CSE DATA SCIENCE

PhishAlert: A Phishing Detector

Tanaya Patil	21107017
Mayank Kumar	21107016
Ayush Mistry	21107029
Sahil Mujumdar	21107050

Project Guide

Prof. Poonam Pangarkar

Outline

- Introduction
- Literature Survey of the existing systems
- Limitations of the existing systems
- Problem statement
- System Design
- Technologies and methodologies
- Implementation
- Conclusion
- References

Introduction

- Motivation :

  - In today's digital landscape, cybersecurity threats pose significant challenges to individuals and organizations worldwide. Phishing attacks, in particular, continue to evolve in sophistication and frequency, targeting unsuspecting users through deceptive emails, websites, and messages.

  - As a result, there is an urgent need for robust and effective mechanisms to detect and mitigate these threats in real-time. The PhishAlert project aims to address this need by developing advanced systems capable of identifying and thwarting phishing attacks, thus enhancing cybersecurity resilience.

Introduction

- Objectives:

- Detect and prevent phishing attacks in real-time by analyzing URL and message characteristics.

- Raise user awareness about phishing risks and provide tools for recognition and prevention.

- Strengthen overall cybersecurity posture by identifying and blocking phishing attempts.

- Drive innovation in cybersecurity through research into new detection and prevention techniques.

Literature Survey of the existing system

- “A Survey of Machine Learning-Based Solutions for Phishing Website Detection” [1] taking into account the paper, the authors find that researchers and security experts have contributed a lot of successful resolutions, from list-based methods and rule-based strategies to machine learning-based approaches. Various machine learning-based solutions achieved higher than 95% accuracy, which is a significant advancement. However, it is believed that the accuracy performance still has space for improvement. The **shortcoming is that this needs an extra feature extraction process** based on rules, and it depends on some third-party services.
- “Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning” [2], here, the authors find that the **MFPD(Multidimensional Feature Phishing Detection)** approach is effective with high accuracy, low false positive rate, and high detection speed. Future development of their approach will consider applying deep learning to feature extraction of webpage code and webpage text. In addition, they plan to implement their approach into a plugin for embedding in a Web browser.

Literature Survey of the existing system

- "Mutual information based logistic regression for phishing URL detection, Cyber Security and Applications" [3] This study investigated how feature selection based on Mutual Information (A useful machine learning technique called Mutual Information (MI) feature selection calculates the statistical dependence of each feature on the target variable) can be **used in Logistic Regression to find fake URLs**. We can see from the results that this method works well for finding a small group of factors that significantly affect the Logistic Regression model's correctness, precision, recall, and F1 score. When MI is used as a feature selection parameter, the Logistic Regression model is easier to understand because it finds the most useful features.
- "A Novel Phishing Email Detection Algorithm based on Multinomial Naive Bayes Classifier and Natural Language Processing" [4] This work proposed a new approach to applying NLP and Naïve Bayes' classifier to detect phishing emails. The **Naïve Bayes classifier is reasonably robust, fast, and accurate.** Most importantly, it is not sensitive to irrelevant features which makes it a suitable choice for email classification. Based on the achieved results, the classifier offers higher accuracy than other works that used the same dataset in the literature which is 97.21%.

Limitations of existing systems

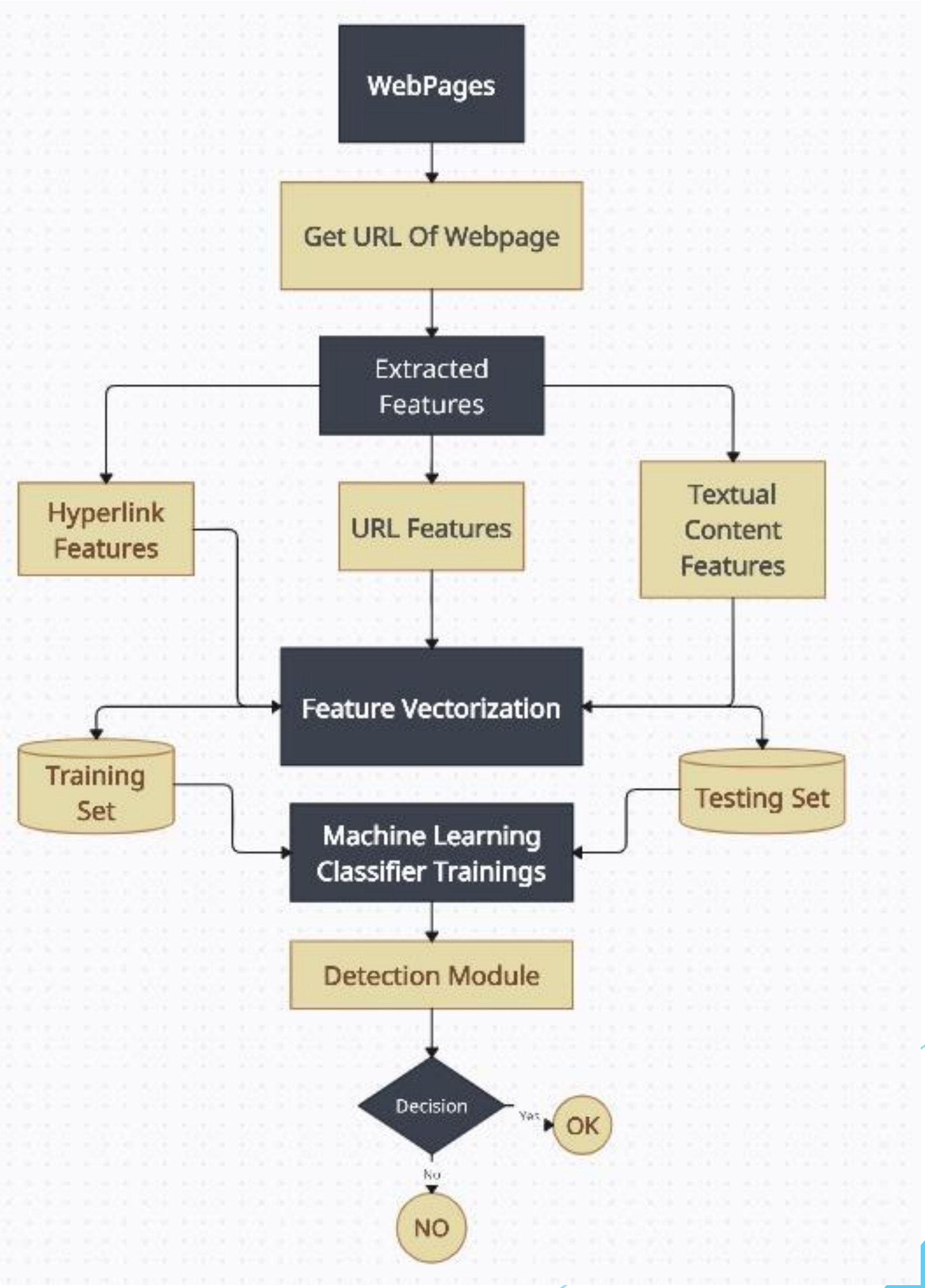
From the literature review of existing systems, we find that,

- The first thing we identified is Multidimensional Feature Phishing Detection is the correct way to approach this problem.
- Various machine learning algorithms such as Logistic Regression, Multinomial Naïve Bayes, AdaBoost Classifier, Bagging Classifier, Extra Trees Classifier, and Gradient Boosting Classifier are tried and tested to find the phishing websites.
- We found out that Logistic Regression and Multinomial Naïve Bayes are the algorithms that work best for the identification of phishing websites and SMS messages and we plan on implementing the same.

Problem Statement

- Phishing is a form of cyber-attack, which hurts people where the user is directed to fake websites and duped to reveal their sensitive and personal information which includes passwords of accounts, bank details, ATM pin-card details, etc. Hence protecting sensitive information from malware or web phishing is difficult.
- This project examines the applicability of ML techniques in identifying phishing attacks and reports their positives and negatives. We will design a Phishing Detection System, that extracts features that are meant to defeat common phishing detection approaches.
- We also make use of classical machine learning techniques like Random Forest, Decision Tree, and wrapper-based features selection which contains the metadata of URLs, and use the information to determine if a website is legitimate or not.

System Design



System Design

Data Collection: Gather email messages, URLs, and metadata from various sources.

Feature Extraction: Convert text data into numerical representations using techniques like tokenization and TF-IDF.

Model Training: Train machine learning models, such as Multinomial Naive Bayes and Logistic Regression, using labeled datasets.

Real-time Monitoring: Deploy trained models for continuous analysis of incoming emails and URLs.

Technologies and Methodologies

- **Logistic Regression:**

1. Data Collection: Gather labeled data.

2. Model Training: Fit a logistic regression model.

3. Prediction: Predict class labels for new data.

4. Decision Boundary: Determine the separating boundary.

5. Evaluation: Assess model performance.

6. Interpretability: Understand feature impact.

- **Multinomial Naive Bayes:**

1. Data Collection: Collect labeled text data.

2. Text Preprocessing: Tokenize and extract features.

3. Model Training: Train a Naive Bayes classifier.

4. Prediction: Predict class labels for new text.

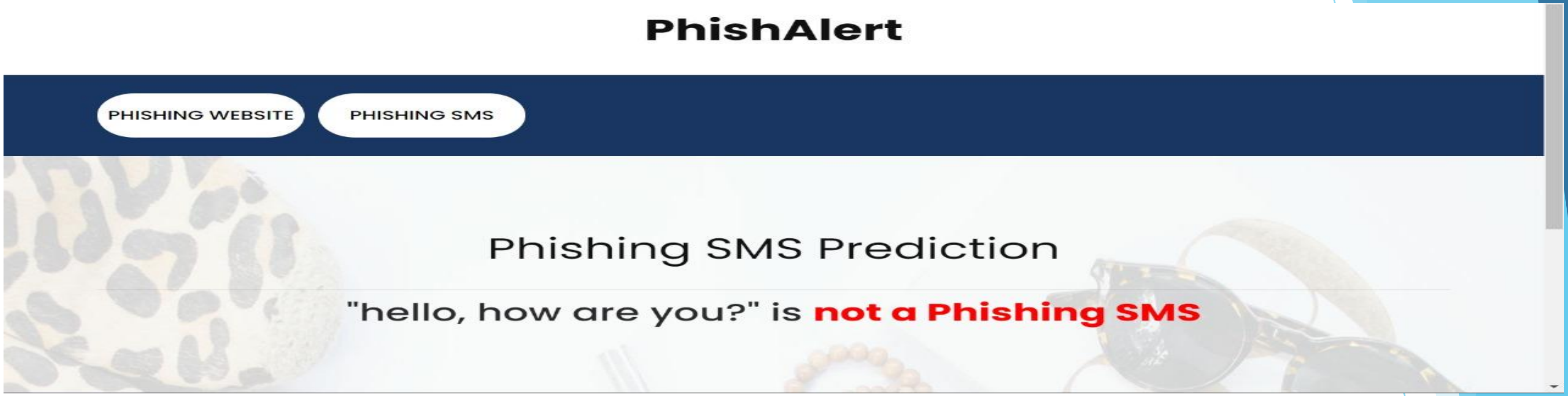
5. Evaluation: Evaluate classifier performance.

6. Interpretability: Analyze class probabilities.

Technologies and Methodologies

- Python: Primary programming language for data preprocessing, modeling, and deployment.
- FastAPI: Web framework for building RESTful APIs to interact with the phishing detection system.
- Scikit-learn: Python library for machine learning tasks, including model training and evaluation.
- NLTK (Natural Language Toolkit): Library for natural language processing tasks, such as text preprocessing and feature extraction.
- Selenium: Web automation tool used for web scraping and data collection.

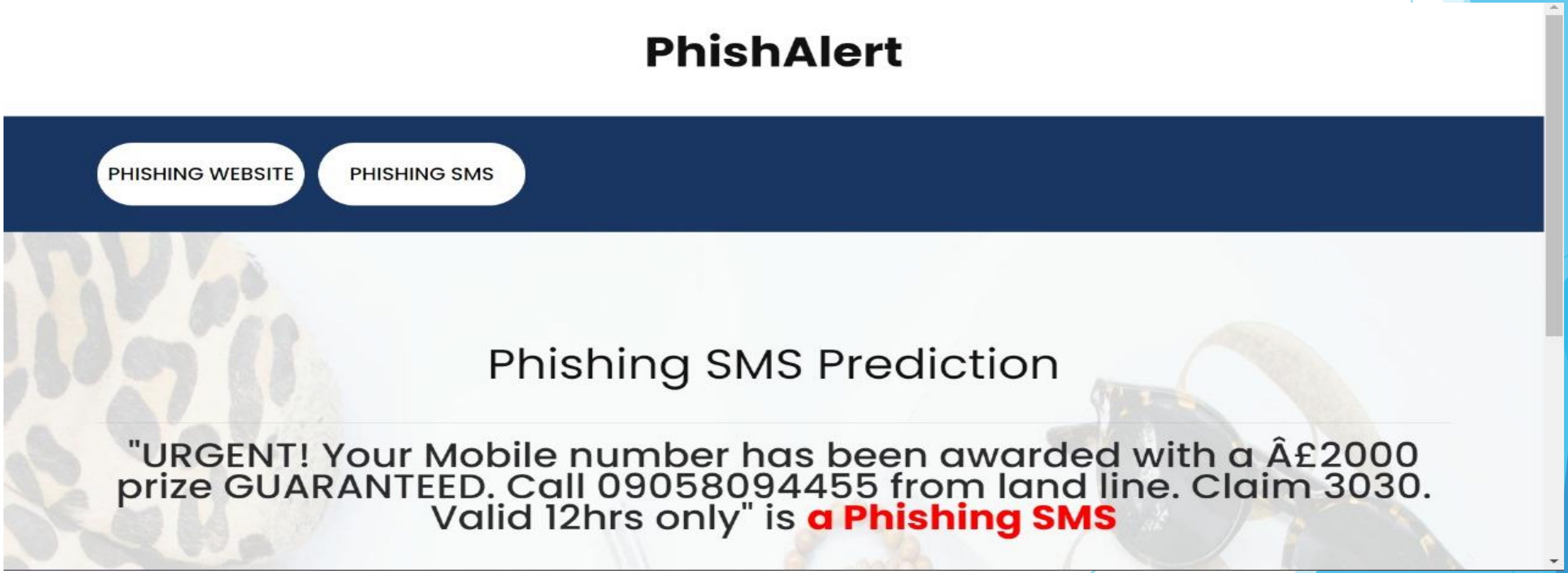
Implementation



PhishAlert

Phishing SMS Prediction

"hello, how are you?" is **not a Phishing SMS**

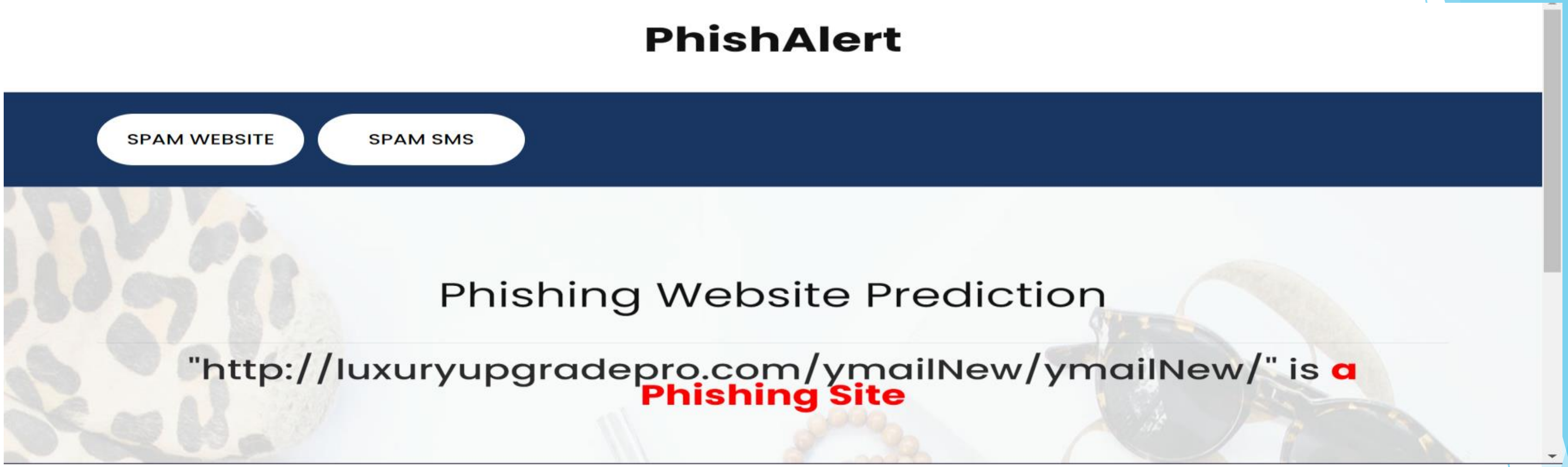


PhishAlert

Phishing SMS Prediction

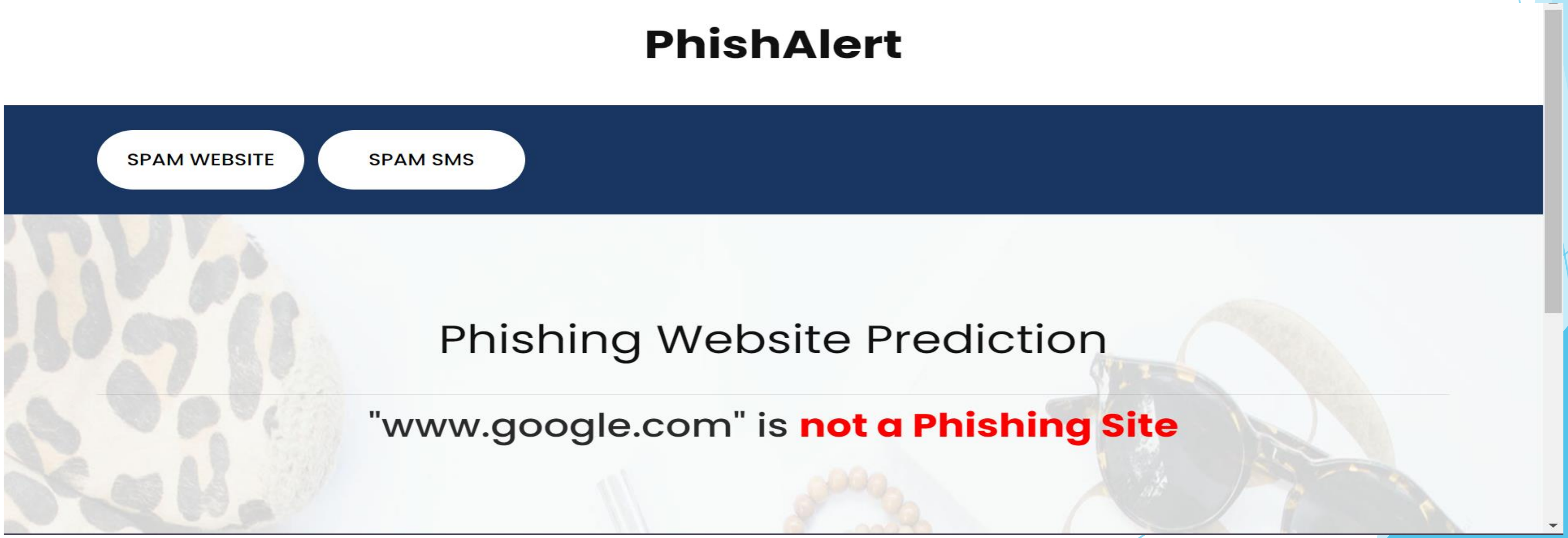
"URGENT! Your Mobile number has been awarded with a Â£2000 prize GUARANTEED. Call 09058094455 from land line. Claim 3030. Valid 12hrs only" is **a Phishing SMS**

Implementation



Phishing Website Prediction

"http://luxuryupgradepro.com/ymailNew/ymailNew/" is **a Phishing Site**



Phishing Website Prediction

"www.google.com" is **not a Phishing Site**

Conclusion

PhishAlert provides an effective solution for detecting and preventing phishing attacks in email and SMS messages.Importance of continuous improvement and adaptation to evolving threats.Future prospects and potential enhancements for further strengthening cybersecurity measures.

References

[1] Tang L, Mahmoud QH. “A Survey of Machine Learning-Based Solutions for Phishing Website Detection”. *Machine Learning and Knowledge Extraction*. 2021; 3(3):672-694.

https://www.mdpi.com/2504-4990/3/3/34

[2] P. Yang, G. Zhao and P. Zeng, “Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning”, in IEEE Access, vol. 7, pp. 15196-15209, 2019, doi: 10.1109/ACCESS.2019.2892066.

https://ieeexplore.ieee.org/abstract/document/8610190

[3] Vajratiya Vajrobol, Brij B. Gupta, Akshat Gaurav, "Mutual information based logistic regression for phishing URL detection, Cyber Security and Applications" 2024, 100044, ISSN2772-9184,

https://www.sciencedirect.com/science/article/pii/S2772918424000109#:~:text=The%20Logistic%20Regression%20model%20is,to%20assess%20the%20model's%20performance

[4] Omar Abdelaziz, Sahana Deb, Rania Hodhod and Lydia Ray “A Novel Phishing Email Detection Algorithm based on Multinomial Naive Bayes Classifier and Natural Language Processing”

Thank You...!!