BITF22M043-Laiba Saleem *Date: November 16, 2025*

Phishing Website Detection System – One Pager

Idea & Problem Statement

This project aims to develop a **real-time phishing detection system** that analyzes URLs using machine learning to protect users from malicious websites. Phishing attacks are becoming increasingly sophisticated, often bypassing traditional security measures, putting users at risk of data theft and financial loss. The solution classifies URLs as **legitimate or phishing with ~97% accuracy**, accessible via Chrome (browser) extension.

Project Summary

Technologies Used: Python, Flask (backend), Logistic Regression & Naive Bayes (ML), Word2Vec & RoBERTa (advanced models), HTML/CSS, Javascript (frontend), pandas, numpy, NLTK (data processing), Pickle (.pkl) for model persistence.

System Workflow:

1. User inputs URL via browser extension.
2. Preprocessing: remove protocol, lowercase, tokenize, stem.
3. Feature extraction using CountVectorizer or embeddings.
4. ML model predicts phishing vs. legitimate.
5. Result displayed on UI/extension (safe / ⚠ phishing).

Dataset: 549,405 URLs (392,983 legitimate, 156,422 phishing).

Achievements & Components

- **ML Models:** Logistic Regression (97% accuracy), Naive Bayes, Word2Vec & RoBERTa.
- **Chrome Extension:** Checks current page or manual input; communicates with Flask API.
- **URL Handling:** Filters browser protocols (chrome://, about:) and internal URLs (localhost, 127.0.0.1).
- **Deployment:** Pickle serialization for production.

Conclusion

This project successfully combines machine learning and natural language processing to deliver a robust phishing detection system. With its high accuracy, real-time performance, and dual interface, the system is a practical solution for combating phishing threats