AI EMAIL PHISHING DETECTOR

A Machine Learning-Powered Solution to Detect and Prevent Email-Based Phishing Attacks

TEAM MEMBERS

Bhargav Raj Dutta - Software Coordinator
Taha Nagdawala - Software Coordinator
Saihan Shafique - Project Activity Coordinator
Roudah Ashfaq - Project Activity Coordinator

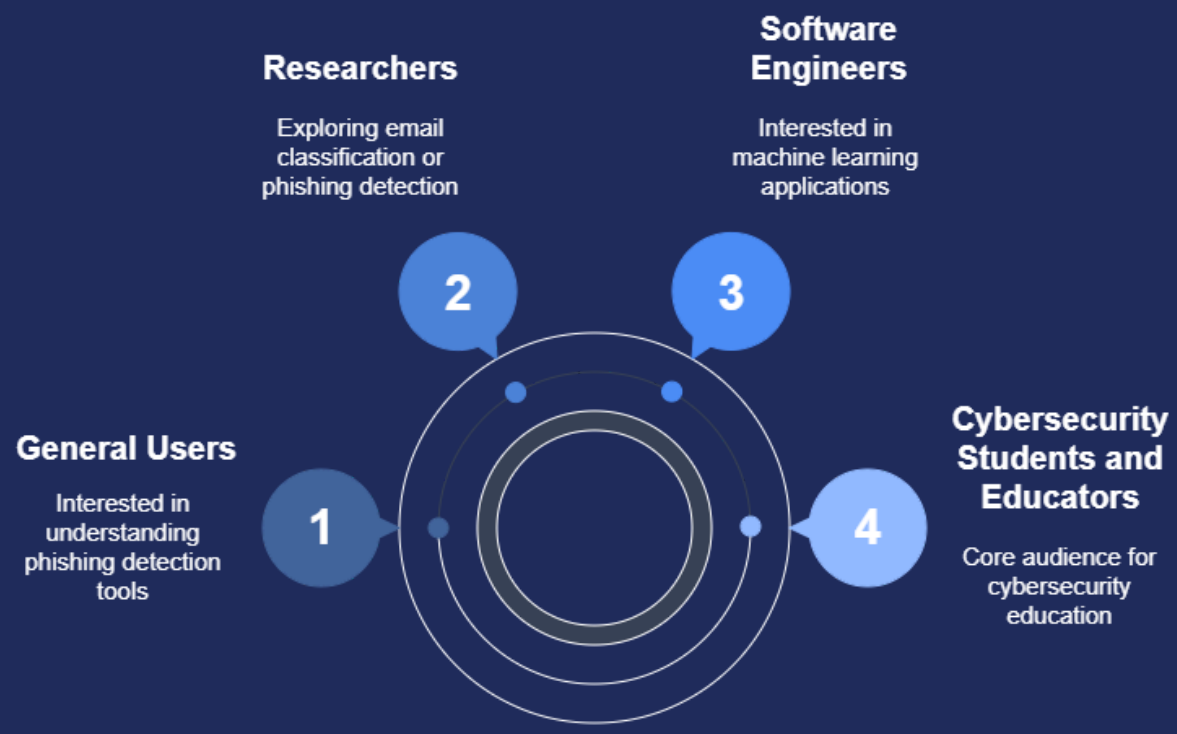
Introduction

This report documents designing, developing, and implementing an AI-based software solution for detecting phishing emails using natural language processing (NLP) and machine learning. As phishing attacks grow in sophistication, there is an urgent need for reliable, automated tools to help users identify potentially malicious emails with high accuracy. This software uses a trained machine learning model and a user-friendly graphical interface to classify email content as **Phishing or Legitimate**.

Purpose of the Report

This report aims to provide technical insight into the architecture, functionality, and implementation of the phishing detector software. It serves as both a development reference and a guide for users and evaluators who wish to understand the core mechanics and objectives behind the tool.

Target Audience



Key Objectives

- Detect phishing emails using a trained AI model.
- Provide a simple and responsive GUI for non-technical users.
- Offer real-time prediction with confidence scores.
- Allow users to test, analyse, and save results for future reference.
- Provide fallback demo mode when the model is unavailable.

Environment

Programming Language

Python 3.x is required for execution.

Machine Learning

Scikit-learn is used for phishing detection.

Vectorization

CountVectorizer converts text to numerical data.

GUI Library

Tkinter is used for the graphical interface.

NLP Tools

NLTK is used for text preprocessing tasks.

OS Compatibility

The software runs on multiple operating systems.

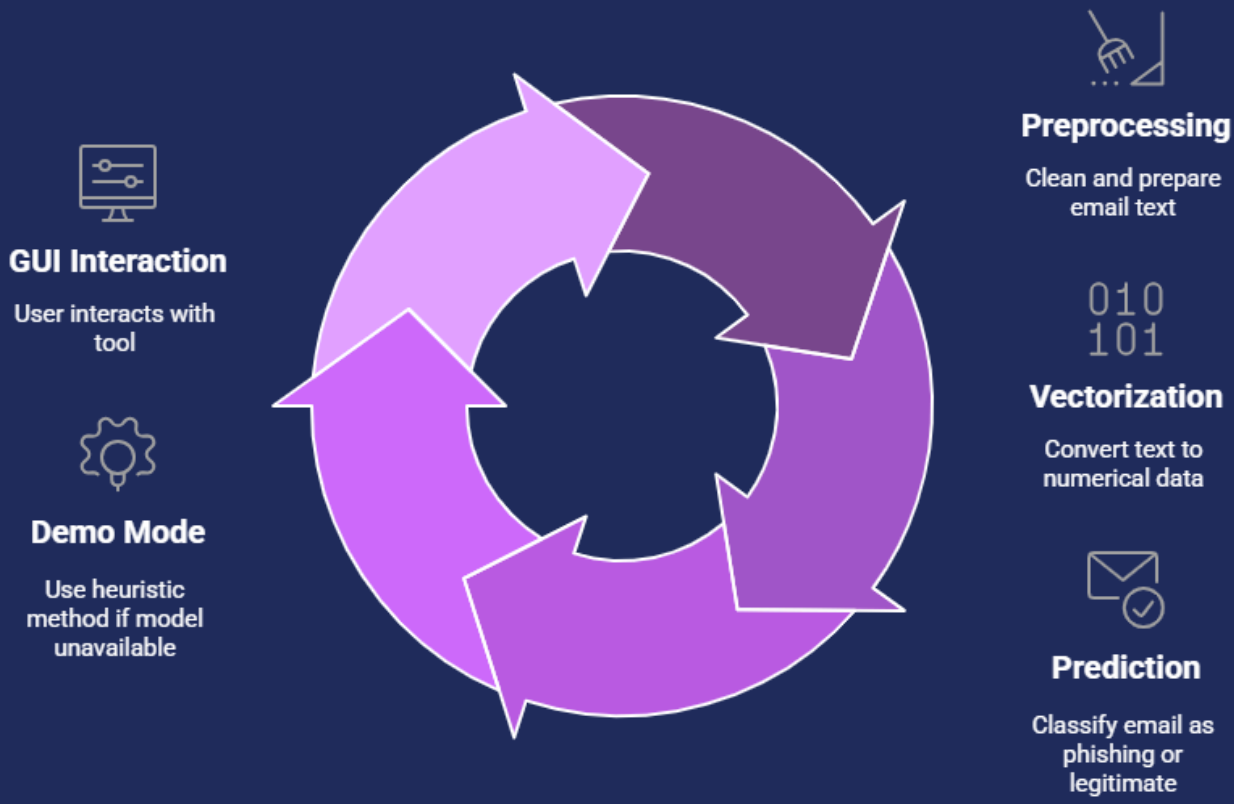
Dependency & Library

- Core libraries: numpy, pandas, re, string, sklearn, joblib, tkinter
- ML & NLP: scikit-learn, transformers, datasets, torch
- Model persistence: pickle, joblib

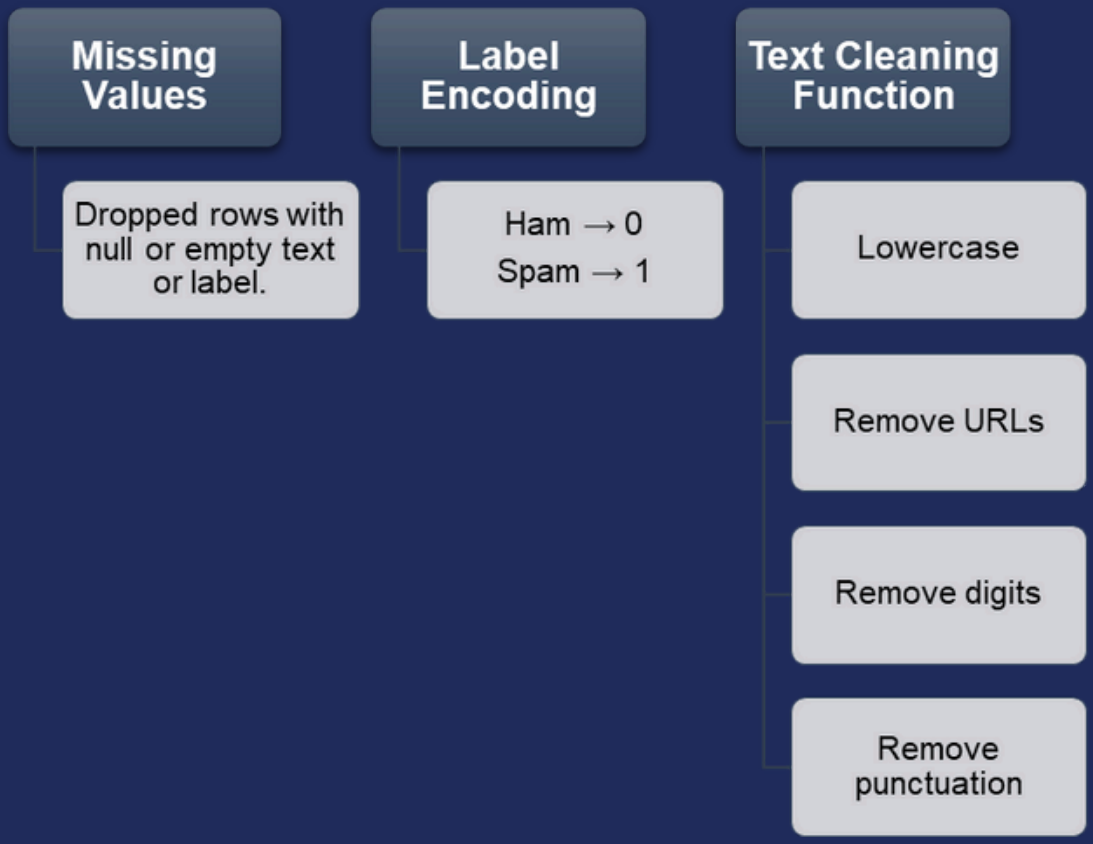
Working of Software

The software uses a machine learning classification model trained on labelled email content to predict whether an email is phishing or legitimate. Here are the major components:

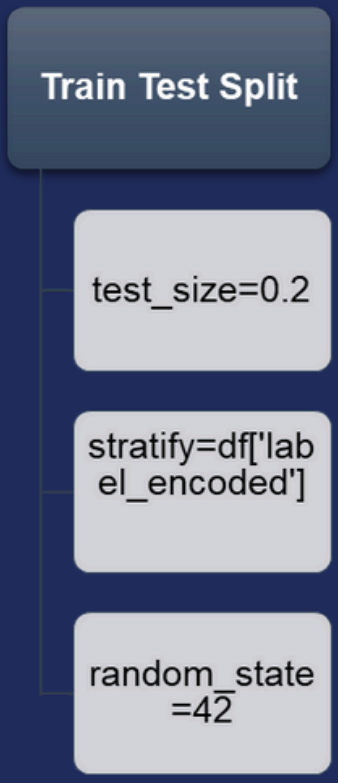
- Preprocessing: The email text is cleaned using lowercase conversion, punctuation removal, number filtering, stopword removal, and tokenisation.
- Vectorisation: The cleaned text is transformed into numerical data using a CountVectorizer.
- Prediction: A pre-trained model classifies the vectorised input and outputs both the label and the probability/confidence.
- Demo Mode: If the model or vectorizer is unavailable, a keyword-based heuristic method is used for prediction.
- GUI: The interface allows users to input emails, view results, and interact with the tool through buttons like Analyse, Paste, Load Sample, and Save Results.



- **Preprocessing and Cleaning**



- **Data Splitting**



- **Baseline Model - Logistic Regression**

1. Vectorizer: TfidfVectorizer with stop_words='english', max_features=5000
2. Classifier: LogisticRegression()
3. Model saved using JobLib

- **Evalutation**

Accuracy

Measures the proportion of correct predictions.

Classification report

Provides precision, recall, and F1-score for each class.

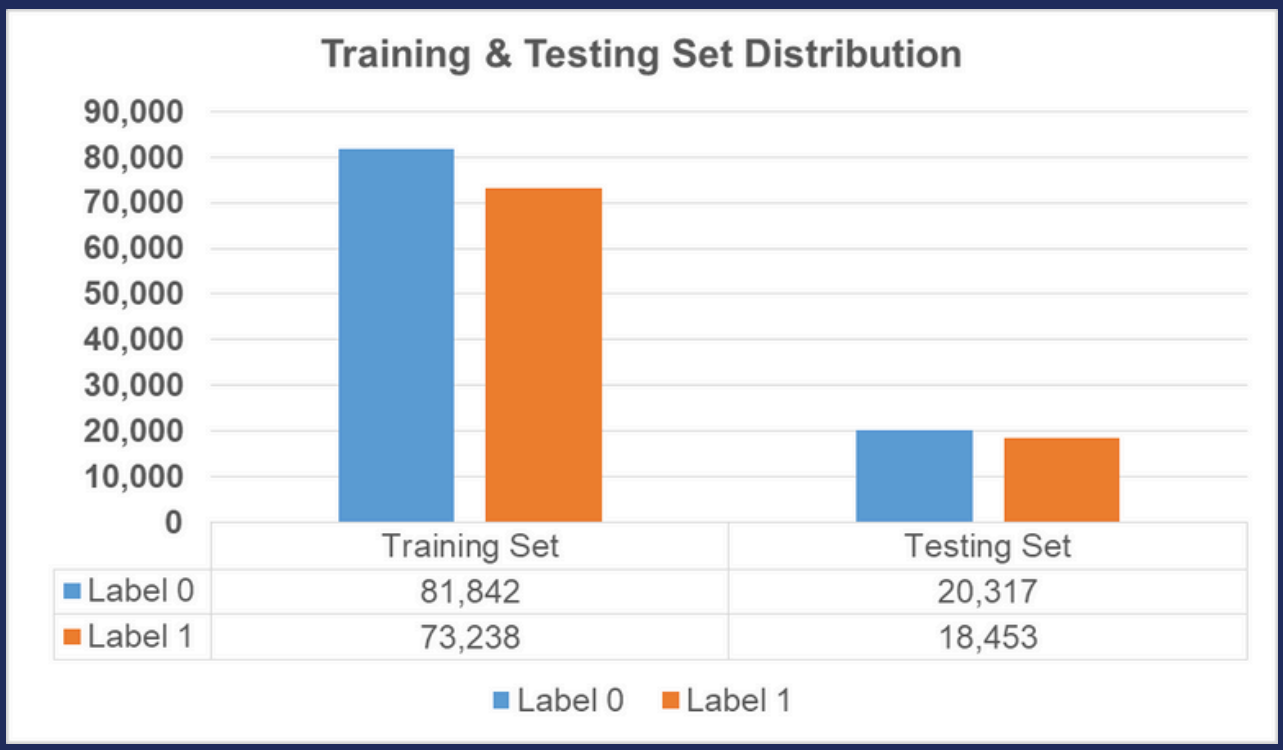
Confusion matrix

Visualizes the performance of a classification model.

- **Data Set Overview**

Attribute	Value
Features Shape	193,850
Label Shape	193,850
Training Set Size	155,080
Testing Set Size	38,770

- **Label Distribution**



- **Data Set Structure**

Dataset Columns	
Text	Contains the content of the email
Label	‘Spam’ or ‘Ham’ for classification

- **Model Performance Metrics**

Model Accuracy

✅ **Model Accuracy: 97.18%**

Classification Report

Class	Precision	Recall	F1 Score	Support
Ham	0.98	0.97	0.97	20317
Spam	0.97	0.97	0.97	18453
Accuracy	-		0.97	38770
Macro Avg	0.97	0.97	0.97	38770
Weighted Avg	0.97	0.97	0.97	38770

- **Confusion Matrix**

	Predicted Ham	Predicted Spam
Actual Ham	19709 (TN)	608 (FP)
Actual Spam	487 (FN)	17966 (TP)

1. *True Negatives (Ham correctly classified): 19,709*
2. *False Positives (Ham classified as Spam): 608*
3. *False Negatives (Spam classified as Ham): 487*
4. *True Positives (Spam correctly classified): 17,966*

Inspiration of work

This project was inspired by the increasing volume and sophistication of phishing attacks that target individuals and organisations via email. Our team recognised the need for a solution that combines the power of machine learning with a user-friendly interface to help detect such threats efficiently.

We researched existing phishing detection techniques, reviewed multiple academic research papers, and utilised publicly available websites that provide dummy phishing and legitimate emails to train and validate our model. The project followed the Agile development process, allowing us to iteratively design, implement, and refine the software based on testing and feedback.

Our goal was to deliver an innovative yet accessible tool leveraging machine learning to contribute to cybersecurity awareness and protection.

Conclusion & Future Work

The AI Email Phishing Detector demonstrates how natural language processing and machine learning can be combined to address a pressing cybersecurity concern. Our team has built a functional prototype that achieves over 97% accuracy in distinguishing phishing from legitimate emails through careful data curation, model training, and interface development.

We aim to expand the dataset for broader generalisation, experiment with deep learning models such as BERT for enhanced accuracy, and integrate our software with existing email platforms for real-time phishing detection. A valuable enhancement would include a link verification mechanism that analyses URLs embedded within emails to detect suspicious or malicious domains. We also envision a web-based version of the tool for increased accessibility.

We hope this project contributes meaningfully to phishing awareness and is a stepping stone for more advanced AI-driven security tools.