EMAIL PHISHING DETECTION USING MACHINE LEARNING

Submitted by: Presented by:-
Abhijit Bhekare
Yash Jangali
Rushikesh Mahadik

Department of Information Technology
Sathaye College Academic Year: 2025–2026

ABSTRACT:

Email phishing is one of the most common and dangerous cyber threats in today's digital world. Phishing emails attempt to deceive users by impersonating trusted organizations in order to steal sensitive information such as login credentials, banking details, or personal data. With the rapid growth of online communication, traditional rule-based email filtering techniques have become insufficient to handle evolving phishing attacks. Hence, there is a need for an intelligent and automated detection mechanism.

This research paper presents an Email Phishing Detection system using Machine Learning techniques. The proposed system analyzes email content and classifies it as phishing or legitimate. A labeled dataset containing phishing and legitimate emails is used for training and testing the model. Text preprocessing techniques such as tokenization, stop-word removal, and vectorization are applied to transform email content into a machine-

readable format. The classification model is trained using supervised machine learning algorithms to improve detection accuracy.
The system demonstrates effective identification of phishing emails with minimal false positives. The results indicate that machine learning-based detection is faster, more adaptive, and more reliable than traditional approaches. This project highlights the practical application of cybersecurity and machine learning concepts and emphasizes the importance of automated phishing detection systems in protecting users from cyber fraud. The proposed solution can be further extended for real-time email filtering and enterprise-level deployment.

1. PROBLEM STATEMENT & OBJECTIVE

Problem Statement

Phishing emails have increased significantly, causing financial loss, identity theft, and data breaches. Users often fail to distinguish between legitimate and phishing emails due to sophisticated attack techniques. Existing traditional filters are not adaptive and fail against new phishing patterns.

Objectives

To detect phishing emails using machine learning

To classify emails as phishing or legitimate

To reduce human dependency in email verification

To improve email security and user awareness

2. LITERATURE REVIEW

Various phishingdetection techniques have been proposed in previous studies. Early methods relied on rule-based filtering and keyword matching, which were limited in accuracy. Later approaches introduced machine learning models such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees. These models showed improved performance but required effective preprocessing and feature extraction. Recent studies emphasize supervised learning and text-based analysis to enhance detection accuracy. However, challenges such as evolving phishing patterns and dataset imbalance remain areas of concern.

3. RESEARCH METHODOLOGY

The methodology followed in this project includes:
Dataset collection
Data preprocessing
Feature extraction
Model training
Email classification
Performance evaluation
A structured workflow ensures accurate and efficient phishing detection.

4. TOOL IMPLEMENTATION

Programming Language: Python
Libraries: Pandas, NumPy, Scikit-learn
Development Tool: VS Code

Dataset: emails.csv
Modules
preprocess.py: Cleans and prepares email text
model.py: Trains the ML model
phishing_detection.py: Takes user input and predicts email type

5. RESULTS & OBSERVATIONS

Model trained successfully
Accurately identifies phishing emails
Fast prediction response
Reduces false positives compared to traditional methods
The system effectively demonstrates the use of machine learning in cybersecurity.

6.ETHICALIMPACT&MARKETRELEVANCE

Ethical Impact
Protects user privacy
Prevents misuse of sensitive data
Promotes responsible AI usage
Market Relevance
Useful for banks and financial institutions
Can be integrated into email services
Helps organizations prevent cyber fraud

7. FUTURE SCOPE

Integration with real-time email systems
Use of deep learning models
Multilingual phishing detection

Browser and email client plugins

8. CONCLUSION

Email phishing is a serious cybersecurity threat. The proposed machine learning-based phishing detection system provides an efficient and reliable solution. The project successfully demonstrates how data preprocessing and supervised learning can enhance email security. This system can be further extended for real-world applications.

9. REFERENCES

APWG Phishing Activity Report
Scikit-learn Documentation
Pandas Official Documentation
Machine Learning by Tom Mitchell
Cybersecurity and Phishing Attacks – IEEE
Python for Data Analysis – Wes McKinney
Email Spam Detection Research Papers
OWASP Phishing Guidelines
Google Safe Browsing Reports
National Cyber Security Centre Publications