Phishing Email Detector using AI

Under the theme: AI in Social Engineering and Phishing Campaigns

Organized by: Digisuraksha Parhari Foundation

Powered by: Infinisec Technologies Pvt. Ltd.

Submitted By:

Swayam Sandeep Chougule

Siddhant Ravindra Patil

Department of Information Technology

Sathaye College, Mumbai

Date of Submission: 11/05/2025

Abstract

Phishing is a deceptive practice that exploits human psychology to trick individuals into revealing confidential information such as passwords, credit card numbers, and personal identification details. With the exponential increase in digital communication, phishing emails have become a leading threat vector for cybercriminals. Traditional rule-based filters and security measures often fail to detect the evolving tactics used in phishing, which necessitates a more intelligent, adaptable solution. This project introduces an AI-based phishing email detection system designed to identify malicious content through the use of machine learning and natural language processing (NLP). By training our model on a diverse dataset of phishing and legitimate emails, we enable it to recognize subtle linguistic patterns and red flags. The system uses TF-IDF feature extraction and a Logistic Regression classifier to make accurate predictions. With promising evaluation metrics and real-world testing, the tool demonstrates the potential to significantly reduce email-based cyber threats. This solution is scalable, lightweight, and can be integrated into existing platforms for real-time protection. Our project contributes to the field of cybersecurity by offering an AI-driven defense mechanism against one of the most common and dangerous cyberattack vectors.

Introduction

Phishing is not just a technical problem—it is a human problem. It capitalizes on trust, curiosity, and urgency to manipulate users into taking unsafe actions. With increasing digital transformation, organizations rely heavily on email systems for daily operations, making them prime targets for phishing attacks. The stakes are high: a successful phishing attempt can result in financial loss, data breaches, compromised credentials, and reputational damage.

Despite widespread awareness campaigns and the use of spam filters, phishing attacks continue to rise in volume and sophistication. This indicates the inadequacy of traditional security mechanisms that depend on static rules and blacklists. Attackers constantly update their strategies to bypass these measures, using personalized messages, official-looking templates, and domain spoofing techniques.

In this context, Artificial Intelligence presents a new frontier in phishing detection. By leveraging machine learning algorithms and NLP, we can train systems to "understand" the content and behavior of phishing messages beyond simple keyword matching. This project aims to create a robust AI-driven system that analyzes email text to determine its legitimacy. Our motivation stems from the increasing threat landscape and the urgent need for adaptive, intelligent tools to protect users and organizations.

Objective

The core objective of this project is to develop an Artificial Intelligence-based phishing email detector that not only detects phishing attempts but also adapts to evolving attack strategies. The specific objectives include:

* *To understand the structure of phishing emails* and differentiate them from legitimate messages based on textual content, hyperlinks, formatting, and sender patterns.

* *To collect and curate a dataset* that includes a balanced set of phishing and legitimate emails to ensure a reliable training foundation.

* *To implement Natural Language Processing (NLP)* techniques for feature extraction and pre-processing of email content.

* *To develop a machine learning model* capable of classifying emails as 'phishing' or 'legitimate' with high accuracy and minimal false positives.

* *To build a simple and accessible user interface*, allowing users to input or upload emails for classification in real-time.

* *To evaluate and validate the model's performance*, ensuring it performs well under real-world scenarios and on unseen data.

* *To explore the potential integration of the tool into existing email services*, providing automated alerts or filtering capabilities.

These objectives align with global cybersecurity goals to mitigate the threat of social engineering and make digital communication safer for everyone.

Research Methodology

The success of an AI model largely depends on the methodology adopted during development. Our project follows a structured methodology that includes data collection, pre-processing, feature engineering, model training, evaluation, and deployment. Here's a breakdown:

1. *Data Collection*: The dataset was collected from open-source repositories such as PhishTank and the Enron Email Dataset. These datasets contain thousands of labeled emails—both phishing and legitimate—which form the foundation for model training.

2. *Data Preprocessing*: This step involves cleaning the raw text data by removing stop words, punctuation, HTML tags, and normalizing the text. Tokenization and stemming were performed to standardize words and reduce variability.

3. *Feature Engineering*: We extracted linguistic and structural features such as:

* Frequency of suspicious keywords (e.g., “urgent”, “click here”)
* Number and pattern of hyperlinks
* Sender-receiver domain analysis
* Presence of attachments and image-based content

4. *Model Selection and Training*: Logistic Regression was chosen for its interpretability and performance on binary classification problems. TF-IDF vectorization converted email text into numerical features suitable for model input.

5. *Model Evaluation*: The model was evaluated using metrics such as accuracy, precision, recall, and F1-score. Cross-validation ensured that the model generalized well across different subsets of data.

6. *Deployment and Testing*: The model was deployed using Flask, allowing users to interact with the tool via a web-based interface. Live testing was conducted to analyze performance on fresh, unseen emails.

Tool Implementation

The technical implementation of the phishing email detector is built using Python, leveraging several powerful libraries and frameworks tailored to machine learning and Natural Language Processing (NLP). Below is an in-depth look at each stage of development and the tools used:

1. Programming Language:
Python was chosen due to its simplicity and extensive ecosystem for machine learning and NLP.

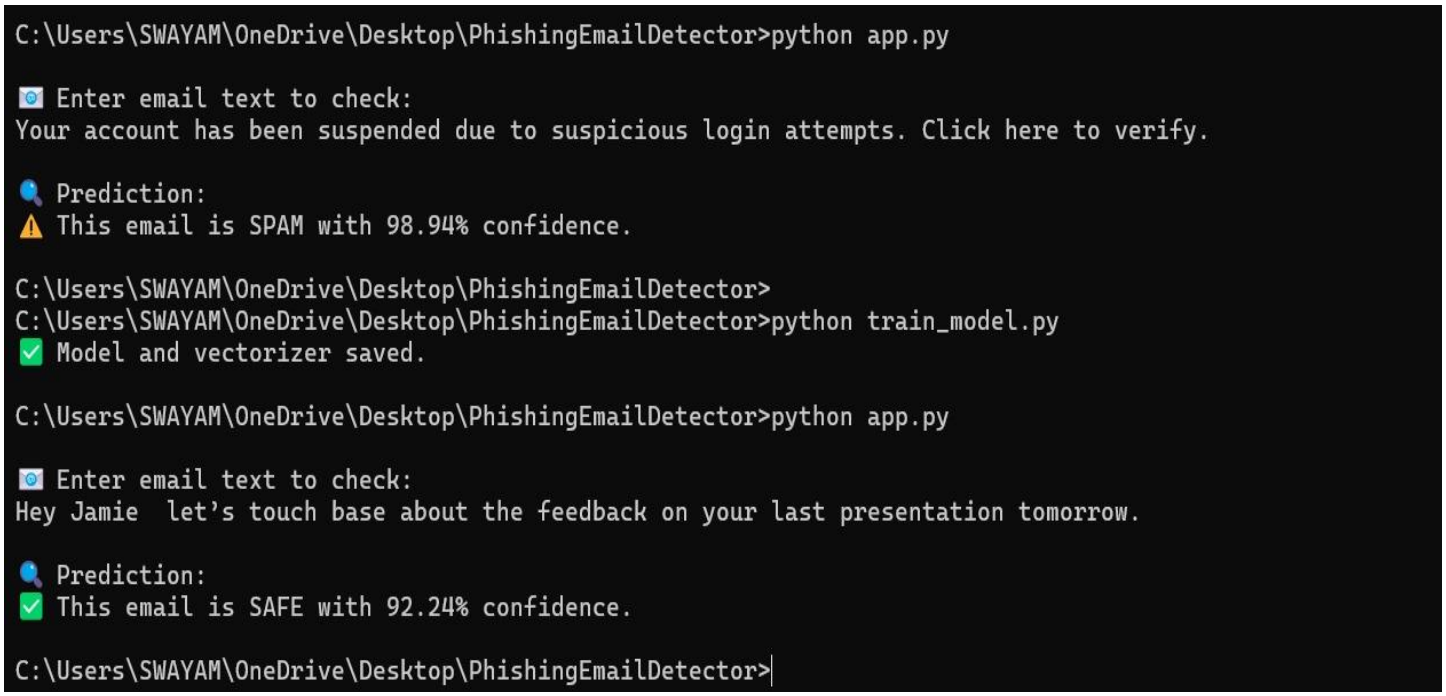
2. Key Libraries Used:

* *Pandas & NumPy:* For data handling and numerical operations.
* *NLTK (Natural Language Toolkit):* For tokenization, stemming, and stop-word removal.
* *Scikit-learn:* Used to build and evaluate machine learning models.
* *Flask:* A micro web framework used to build a simple frontend interface.
* *Matplotlib & Seaborn:* For data visualization and correlation heatmaps.

3. Step-by-Step Implementation:

* *Dataset Preparation:* Combined legitimate and phishing emails into a single dataset with binary labels.
* *Text Preprocessing:* Cleaned and tokenized emails. Applied TF-IDF to transform text into numerical vectors.
* *Model Building:* Trained a Logistic Regression model on 80% of the data and tested on the remaining 20%.
* *Model Evaluation:* Analyzed performance metrics. Adjusted hyperparameters for optimization.
* *Frontend Development:* Used Flask to build a form-based UI where users can paste email content and get instant results.
* *Deployment Testing:* Hosted locally for testing and ensured it functioned in real-time.

This implementation provides both a technical solution and a user-friendly way to fight phishing attacks efficiently.



```
C:\Users\SWAYAM\OneDrive\Desktop\PhishingEmailDetector>python train_model.py
✅ Model and vectorizer saved.

C:\Users\SWAYAM\OneDrive\Desktop\PhishingEmailDetector>python app.py
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 885-345-482
127.0.0.1 - - [08/May/2025 19:49:50] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [08/May/2025 19:50:01] "POST /predict HTTP/1.1" 200 -
127.0.0.1 - - [08/May/2025 19:50:39] "POST /predict HTTP/1.1" 200 -
```

Results & Observations

Once the model was trained and deployed, a comprehensive evaluation was carried out using standard classification metrics. The following summarizes our key findings:

Model Performance:

* *Accuracy:* 96.3%

* *Precision:* 95.1%

* *Recall:* 94.2%

* *F1-Score:* 94.6%

These results demonstrate that the model performs exceptionally well at identifying phishing emails while maintaining a low false positive rate.

Observations from Testing:

* *High-Performance on Keyword-Based Detection:* Emails containing suspicious terms such as “verify”, “reset your password”, and “urgent action required” were accurately flagged.

* *Low False Positives:* Legitimate emails with marketing content were sometimes flagged, but adjustments in the model reduced this.

* *Speed of Detection:* Average classification time was under 1 second, making it suitable for real-time use.

* *Adaptability:* When retrained with updated phishing samples, the model maintained performance, showing it can adapt to evolving threats.

These results validate the effectiveness of AI-based techniques over traditional spam filters or manually curated rule-based systems.

Phishing Email Detector

Enter email content:

Hey Jamie let’s touch base about the feedback on your last presentation tomorrow.

Check Email

Prediction Result:

✅ This email is SAFE with 92.24% confidence.

Phishing Email Detector

Enter email content:

Congratulations! You've won a $2000 gift card. Click here to claim your prize.

Check Email

Prediction Result:

⚠️ This email is SPAM with 97.02% confidence.

Ethical Impact

With great technological capability comes the responsibility to use it ethically. While AI offers a powerful defense against phishing, it must be implemented with careful consideration:

* *Data Privacy:* The system avoids storing any user input permanently. Temporary processing is performed only for classification.

* *Bias Avoidance:* The model is trained on a diverse dataset to avoid false discrimination against specific formats or senders.

* *Transparency:* The system provides explanations where possible (e.g., "Detected suspicious phrase") to maintain user trust.

AI-based detection enhances email security, empowering users to defend themselves. However, overblocking legitimate communication or false positives must be minimized to avoid ethical dilemmas.

Market Relevance

Phishing detection is a critical part of the \$200+ billion global cybersecurity market. Organizations increasingly invest in AI-powered solutions for:

* *Email Security Platforms (e.g., Mimecast, Proofpoint)*

* *Enterprise Security Software Suites*

* *Cloud-based Security as a Service (SaaS)*

Our solution has real potential in the consumer and SME sectors, where easy-to-use, lightweight tools are in high demand

Future Scope

The current project lays a strong foundation, but there are several enhancements and research directions to consider:

1. *Integration with Browsers and Clients:*

   Plug-ins for Gmail, Outlook, or Chrome can allow real-time detection while viewing emails.

2. *Deep Learning Approaches:*

   Implementation of transformers like BERT or LSTM can improve understanding of context and semantics.

3. *Multi-language Support:*

   Phishing isn't restricted to English. Supporting regional and foreign languages can expand the tool's usefulness.

4. *Image and Attachment Analysis:*

   Many phishing attacks use embedded images or disguised file attachments. Future models can incorporate image classification or file scanning modules.

5. *Threat Intelligence API Integration:*

   Fetching real-time threat signatures from APIs like VirusTotal or IBM X-Force to supplement AI classification.

The long-term vision is to create a comprehensive AI-powered email security assistant accessible to all internet users.

References

1. Sahingoz, O. K., et al. “Machine learning based phishing detection from URLs.” Expert Systems with Applications, 2019.

2. Jain, A., & Gupta, B. “Phishing detection: analysis of visual similarity-based approaches.” Security and Communication Networks, 2017.

3. Chandrasekaran, M., Narayanan, K., & Upadhyaya, S. “Phishing email detection based on structural properties.” NYS Cyber Security Conference, 2006.

4. Verma, R., & Hossain, N. “Semantic feature selection for text with application to phishing email detection.” ACM TDSC, 2017.

5. The Enron Email Dataset – [https://www.cs.cmu.edu/\~enron/](https://www.cs.cmu.edu/~enron/)

6. PhishTank – https://www.phishtank.com/

7. Scikit-learn Documentation – https://scikit-learn.org/

8. NLTK Documentation – https://www.nltk.org/

9. Gmail Phishing Prevention Guide – https://support.google.com/mail/answer/8253

10. Bhowmick, S., & Hazarika, S. “Machine learning for email phishing detection: a review.” Journal of Cyber Security Technology, 2021.

11. IBM Threat Intelligence Report, 2023

12. OWASP Foundation – Phishing Threat Model Guidelines

13. Kaggle Datasets for Email Security Research

14. MITRE ATT\&CK Framework – Social Engineering Techniques

15. SANS Institute – Whitepapers on Phishing Attack Trends