AI-Powered Phishing Email Detector

By Jordan Chelsey, Elton Batista, Curtis Jones

Advised by Khaled Slhoub

Client: Khaled Slhoub

Client Meetings: Monday at 4pm every 2 weeks, first meeting 9/15

Goal and Motivation:

The goal of this system is to analyze the user's inbox and use an AI model to discern which emails are "phishing" scams. Phishing scams are defined as, "the fraudulent practice of sending emails or other messages purporting to be from reputable companies in order to induce individuals to reveal personal information, such as passwords and credit card numbers". These emails typically target technologically illiterate people, such as children or the elderly. However, anybody can fall for these scams, posing a personal security risk to anybody who owns an email account. Our motivation behind this product lies with the idea that people may be bombarded with these emails with no built-in system from email hosting websites, and the only way to discern these emails is a person's own personal discretion. Our goal is to help minimize risk from an individual's inbox by developing software to detect phishing emails.

Approach:

Email Analysis and Classification Using AI

By checking the subject, the sender, the body text, and any links attached to an email, a trained AI model will be able to classify emails based on risk analysis. Using a dataset of known phishing emails, scams that are not normally flagged by a spam filter will be able to be discerned from legitimate emails using a combination of factors determined to be suspicious by the model. This includes checking for suspicious domain names from the sender, links, and keywords (i.e., Update your account NOW!!!). Emails in an inbox are given a "risk score" based on the model's confidence that a given email is phishing. Emails above a certain threshold are classified as "suspicous", and those near guaranteed by the model to be phishing scams are classified as "phishing email".

Gmail Integration Tools

By logging in through Google, an inbox from any domain can be checked for phishing scams. The AI will be able to sift through and check for patterns in an inbox on top of the analysis of known phishing emails. A variety of functions can be performed on high-risk emails, such as deletion and reporting.

Visualization

Emails with a higher risk are sorted to the top of the page. Each flagged email will be given an explanation as to why the model believes it is a scam. This gives users a better understanding of why an email should be removed from their inbox. Several visual tools, such as charts and diagrams, will allow users to see the percentage of emails in their inbox that are flagged as “phishing emails”, “suspicious”, and “legitimate”. By doing this, users gain a better understanding of how at risk they are to phishing scams on the internet. The model will also be able to give tips to users on internet safety by analyzing possible sources of these emails to avoid falling victim in the future.

Novel Features/Functionality:

Most phishing detectors require pasting individual emails into text boxes on websites. The novelty of this system comes from the integration of AI and Gmail’s API. The model can find patterns within a user's inbox and allow them to perform multiple functions on high-risk emails. An analysis of known phishing emails and a user’s inbox will allow this system to be more accurate than any phishing email detector while allowing for more functionality on what to do with those emails.

Algorithms and tools:

- AI Models
    - Machine Learning
        - Random Forest
        - Transformer
            - DistilBERT
    - LLM
        - Ollama
        - Grok
        - Meta AI
- Gmail API
    - All emails will have to be logged in through a token with Gmail

Technical Challenges:

A large challenge we face is deciding between different AI models. The AI Model we choose will affect both development and end-results for the user at the end of the project.

- Advantages and Disadvantages of different AI Models
    - Machine Learning
        - Random Forest
            - Non-linear model
            - Computationally expensive
            - Strong predictive power in larger data sets
        - Transformer Models
            - Neural network
            - Good with natural language processing
            - Could be more accurate than an LLM
    - LLMs (Large Language Models)
        - Pre-trained to understand existing language patterns
        - Can be highly susceptible to faults in output
        - Can better understand conceptual and abstract sentences

Training our AI model on a quality dataset is extremely important if we want to have accurate measurements on real emails.

- Requirements for good data sets
    - Contains the following components of an email
        - Email of sender

        - Subject
        - Body
    - Includes examples of “phishy” links
        - I.e. shortened URLs
    - Large data set
        - By using a larger data set, we reduce the margin of error in the trained model

In addition to this, some of our members are unfamiliar with web development.

Milestone 1:

Milestone 1 (Sep 29): itemized tasks:

- Resolve technical challenges:
    - Compare and select quality datasets
    - Compare and select an AI model that fits the needs of the project
        - Test different AI models with different datasets to determine which yields the best results
    - Catch members up on JavaScript development
- Create a small demo using the selected AI model
    - I.e. a textbox that scores different emails on how likely they are to be phishing scams
- Compare and select collaboration tools for software development, documents/presentations, communication, task calendar
    - Create discord, Trello, and Github to collaborate on project
- Create Requirement Document
- Create Design Document
- Create Test Plan

Milestone 2:

- AI Model Training
    - Use selected dataset to feed and train selected AI model
        - Recognizing suspicious senders
        - Recognizing suspicious links
        - Recognizing suspicious language patterns (i.e. urgent language)
    - Generate a user-friendly summary of why an email is flagged as suspicious
- AI Model Testing
    - Gather testing data to forward test the detection model. Testing data can be:
        - Real phishing scams from the team’s inboxes
        - Generated phishing scams using AI
        - A reserved portion of the selected dataset
    - Use results of testing to tune detection model
- Research Gmail Integration
    - Look into the services required to use Google’s Gmail API
    - Read API documentation for accessing, reading, and deleting messages from the user’s inbox
    - Research Google OAuth integration for the web interface

Milestone 3:

- Complete Gmail Integration
  - Feed actual user inbox messages directly into the detection model
  - Allow the model to delete suspicious messages when given permission to do so
- Begin Web Interface Construction
  - Implement Google OAuth
  - List the user's messages on the dashboard

- *Task matrix for Milestone 1 (teams with more than one person)*

Task	*Jordan*	*Elton*	*Curtis*
Compare and select Technical Tools	*Datasets*	*Machine Learning Models*	*LLM Models*
Detection Demo	*Web App*	*AI Model Training*	*AI Model Testing*
Resolve Technical Challenges	*Testing Different Data Sets*	*Machine Learning Testing*	*LLM Testing*
Compare and select Collaboration Tools	*Trello setup*	*Discord Setup*	*Github setup*
Requirement Document	*write 33%*	*write 33%*	*write 33%*
Design Document	*write 33%*	*write 33%*	*write 33%*
Test Plan	*write 33%*	*write 33%*	*write 33%*

- *Approval from Faculty Advisor*
    - *"I have discussed with the team and approve this project plan. I will evaluate the progress and assign a grade for each of the three milestones."*
    - *Signature: ______Dr. Khaled Slhoub________________ Date: __9/3/2025______*