Chatbot for Phishing Detection Using NLP and Deep Learning

Author: Roberto J Gonzalez Hernandez
Advisor: Jeffrey Duffany, Ph.D.
Computer Science Department

Abstract

Phishing remains one of the most prevalent and damaging cybersecurity threats, exploiting human trust through deceptive emails, messages, and websites. In a single quarter of 2022, over one million phishing attacks were recorded, responsible for 80% of security incidents and significant financial losses. This project presents an intelligent chatbot that uses Natural Language Processing (NLP), deep learning, and OpenAI's GPT-3.5 to detect phishing attempts in real-time. The chatbot processes both text and images (using OCR) and provides clear, human-readable explanations. Preliminary results show high accuracy, improving user awareness and offering a powerful tool to combat online scams effectively.

Introduction

Phishing is a widespread and evolving cybercrime where attackers impersonate trusted sources to steal sensitive data or spread malware. In 2023, phishing was responsible for 80% of cybersecurity incidents, with over a million attacks recorded in a single quarter. Traditional defenses, such as rule-based filters and blacklists, are increasingly ineffective against sophisticated tactics like AI-generated messages, embedded text in images, and QR code obfuscation. This project presents an intelligent chatbot powered by GPT-3.5, integrating Natural Language Processing and deep learning to detect phishing attempts in real-time. It analyzes both text and image inputs, providing users with clear, informative explanations and helping strengthen their cybersecurity awareness.

Background

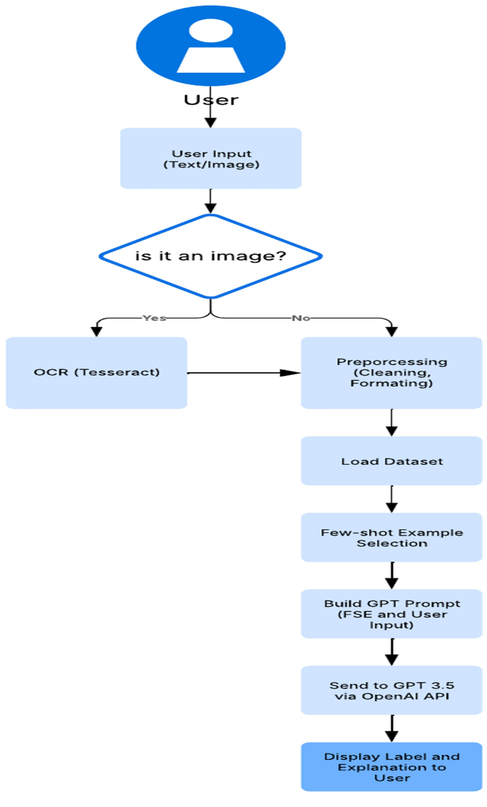
Advancements in cybersecurity have transformed phishing detection from basic rule-based methods to sophisticated AI-driven solutions. Early techniques, such as blacklists and keyword matching, lacked adaptability and struggled with novel or disguised threats. Today, modern approaches combine natural language processing (NLP), computer vision, and optical character recognition (OCR) to analyze content more effectively—including text embedded in images. The development of large language models like GPT-3.5 introduces powerful tools capable of understanding context, tone, and subtle indicators of phishing. This project integrates GPT-3.5 and OCR into an intelligent chatbot that detects phishing attempts in real time, while educating users through clear, human-readable explanations to strengthen awareness and protection.

Problem

Traditional detection systems often fail to identify phishing attempts disguised as legitimate messages, especially those using images, QR codes, or personalized content. Many users lack accessible tools to verify suspicious emails. This project introduces a chatbot powered by GPT-3.5 and OCR to analyze both text and image inputs, providing clear explanations to improve phishing detection, reduce false positives, and enhance user awareness in real time.

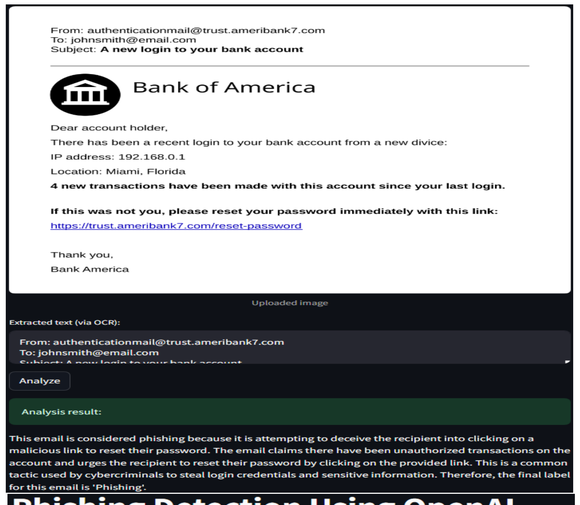
Methodology

To address the challenge of phishing detection, we developed an intelligent chatbot with a streamlined, multi-stage architecture. The system is implemented using Streamlit and allows users to input either text or image files, accommodating a wide range of phishing formats, including screenshots and QR code-based scams. Text inputs undergo minimal preprocessing, while image inputs are processed using Tesseract OCR to extract embedded text for analysis. This is crucial for identifying phishing content hidden in visual formats. The extracted or direct text is then sent to OpenAI's GPT-3.5 model via API, using a structured prompt that instructs it to act as a cybersecurity expert. The model evaluates the message based on sender details, URLs, tone, and content, and returns a classification—"Phishing" or "Legitimate"—along with a clear explanation. The chatbot displays the results in a user-friendly interface, such as: "Classification: Phishing. Warning: This message includes a suspicious link and urges immediate action," or "Classification: Legitimate. No phishing indicators were detected." Users can ask follow-up questions or input additional messages for analysis. The chatbot handles each case independently, though limited conversational context can be retained if necessary. We tested the system using real phishing and legitimate samples, fine-tuning the prompts and validating the OCR accuracy. The application runs locally or in the cloud, responds in 5–8 seconds, and includes basic anonymized logging for ongoing improvement. This combination of OCR, NLP, and GPT-3.5 creates a practical, accessible, and educational tool for real-time phishing detection.



Results and Discussion

The phishing detection chatbot was tested using a mix of phishing and legitimate emails, image-based content, and user-generated inputs to assess its accuracy and usability. In a sample of 20 phishing and 20 legitimate emails, the chatbot correctly identified 95% of phishing cases and 90% of legitimate ones. False positives occurred when messages had an urgent tone or misleading language, but even in those cases, the chatbot provided reasonable explanations. The system excelled in identifying known phishing patterns, such as suspicious URLs, generic greetings, and urgency, mimicking how a human expert would assess a message. Image-based tests also proved successful. The chatbot flagged phishing messages embedded in images, such as fake file share notifications and QR code scams. Although it couldn't decode QR codes directly, it warned users based on surrounding context. Real users found the tool easy to use and appreciated the clear, educational feedback. The explanations helped confirm suspicions and build user awareness. The system performed reliably, with response times of 5–8 seconds per query. However, limitations include reliance on the OpenAI API (raising privacy and availability concerns) and the manual nature of usage—users must choose to consult the chatbot. Still, its educational value and contextual reasoning demonstrate that a GPT-3.5-powered chatbot is a promising tool for real-time phishing detection and user training.


From: authenticationmail@trust.ameribank7.com
To: johnsmith@email.com
Subject: **A new login to your bank account**

Bank of America

Dear account holder,
There has been a recent login to your bank account from a new divice:
IP address: 192.168.0.1
Location: Miami, Florida
4 new transactions have been made with this account since your last login.

If this was not you, please reset your password immediately with this link:
https://trust.ameribank7.com/reset-password

Thank you,
Bank America

Uploaded image

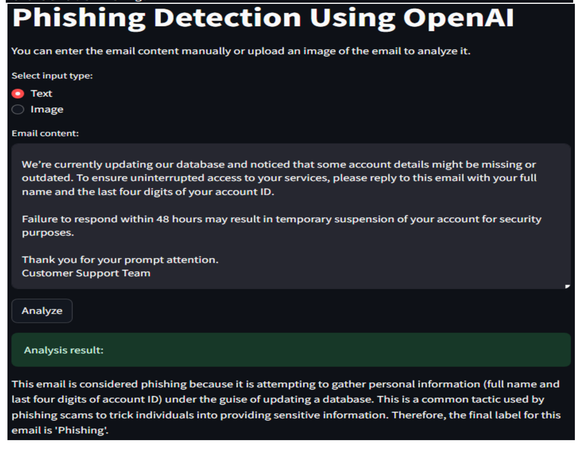
Extracted text (via OCR):

From: authenticationmail@trust.ameribank7.com
To: johnsmith@email.com

Analyze

Analysis result:

This email is considered phishing because it is attempting to deceive the recipient into clicking on a malicious link to reset their password. The email claims there have been unauthorized transactions on the account and urges the recipient to reset their password by clicking on the provided link. This is a common tactic used by cybercriminals to steal login credentials and sensitive information. Therefore, the final label for this email is 'Phishing'.


Phishing Detection Using OpenAI

You can enter the email content manually or upload an image of the email to analyze it.

Select input type:
Text
Image

Email content:

We're currently updating our database and noticed that some account details might be missing or outdated. To ensure uninterrupted access to your services, please reply to this email with your full name and the last four digits of your account ID.

Failure to respond within 48 hours may result in temporary suspension of your account for security purposes.

Thank you for your prompt attention.
Customer Support Team

Analyze

Analysis result:

This email is considered phishing because it is attempting to gather personal information (full name and last four digits of account ID) under the guise of updating a database. This is a common tactic used by phishing scams to trick individuals into providing sensitive information. Therefore, the final label for this email is 'Phishing'.

Conclusions

This project presents an Intelligent Chatbot for Phishing Detection using Natural Language Processing, GPT-3.5, and Optical Character Recognition (OCR) to analyze both text and image-based messages. Unlike traditional filters, the chatbot provides clear explanations alongside its phishing verdict, helping users understand threats through expert-like feedback. Testing showed high accuracy, correctly identifying phishing patterns such as suspicious URLs and urgent requests. Users reported increased awareness and confidence in handling suspicious messages. The chatbot's architecture supports rapid deployment without retraining, making it adaptable to evolving threats. While future improvements are needed for privacy and adversarial protection, the chatbot offers a practical, educational, and accessible cybersecurity tool.

Future Work

Future work includes integrating the chatbot with the Gmail API to enable automatic scanning of incoming emails for phishing detection. This would allow the system to analyze messages in real time without requiring manual input from users, improving usability and response speed. By automating the detection process, the chatbot could proactively alert users to suspicious content within their inboxes, offering both protection and education. This integration would significantly enhance its practicality for everyday cybersecurity use.

Acknowledgements

I would like to sincerely thank Dr. Jeffrey Duffany for his guidance and support throughout the development of this project. His expertise and feedback were essential to the successful completion of this work. This project was independently developed and did not require external funding—only time, dedication, and personal effort.

References

[1] Subhajit Bandyopadhyay, Phishing Emails Dataset, Kaggle, 2021. [Online].Available: https://www.kaggle.com/datasets/subhajournal/phishingem ails
[2] Keepnet Labs, "2025 Phishing Statistics: Top Phishing Stats, Insights & Trends." 2025. [Online]. Available: https://keepnetlabs.com/blog/top-phishing-statistics-andtrends-you-must-know
[3] Anti-Phishing Working Group (APWG), Phishing Activity Trends Report, 1st Quarter 2022, May 2022. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q1_2022. pdf
[4] A. Ejaz, A. N. Mian, and S. Manzoor, "Life-long phishing attack detection using continual learning," Scientific Reports, vol. 13, art. 11488, Jul. 2023.
[5] Securelist (Kaspersky), "Investigating ChatGPT phishing detection capabilities," Feb. 2023. [Online]. Available: https://securelist.com/chatgpt-anti-phishing/109590/
[6] A. Rusk, "Leveraging Artificial Intelligence in the Fight Against Phishing," INKY Email Security Blog, Jul. 2023inky.com. [Online]. Available: https://www.inky.com/en/blog/leveraging-artificialintelligence-in-the-fight-against-phishing
[7] N. Alarcon, "OpenAI Presents GPT-3, a 175 billion Parameters Language Model," NVIDIA Technical Blog, Jul. 7, 2020developer.nvidia.com. [Online]. Available: https://developer.nvidia.com/blog/openai-presents-gpt-3-a175-billion-parameters-language-model/
[8] Smadi, H., Aslam, N., and Zhang, L., "Detection of Online Phishing Email Using Dynamic Evolving Neural Network Based on Reinforcement Learning," Decision Support Systems, vol. 107, pp. 88–102, 2018nature.com. (Cited in Scientific Reports 2023).