COS 720 Project

Zion van Wyk

February 2025

1 Overview

Phishing detection using AI for email-based threats identifies attempts to steal sensitive information, such as login passwords or financial data, by sifting through large amounts of email data. Through pattern recognition, these systems flag suspicious emails in real-time [2]. By using machine learning algorithms, such as decision trees, support vector machines, and neural networks, AI has been used extensively in phishing detection [3]. Natural Language Processing techniques have also been used to identify phishing attempts in email content through linguistic analysis, highlighting suspicious content.

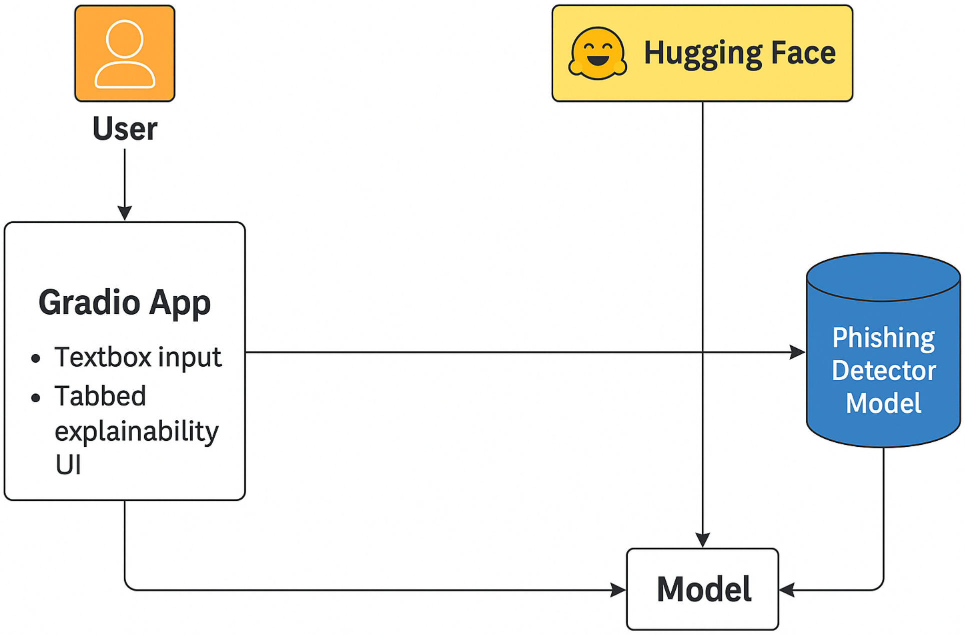
Unlike traditional methods, these AI-powered systems can adapt with the rise of new and advanced methods of phishing with improved accuracy. AI's ability to learn from new data allows it to stay ahead of emerging threats that become increasingly complex. For example, context-aware phishing which uses a target's interests and communication patters to create personalised phishing emails [1]. AI thus provides a significant advantage over rule-based or manual approaches that traditional methods provide.

2 Model Selection

2.1 Selection Process

A robust but lightweight model was needed for this project, due to the limitations in computational power. In true, modern fashion, a model that follows the Transformer architecture was chosen; DistilBERT. BERT has excellent classification capabilities but DistilBERT is 40% smaller yet 60% faster with 97% of BERT's original capabilities. [4]

2.2 System Architecture & Workflows



2.3 Potential Limitations

- Overfitting is possible if the dataset is not diverse and doesn't cover a range of phishing tactics, like spear-phishing, or overly sophisticated emails impersonating close contacts. The model may not generalise well because of this.

- Since this model is smaller, it may miss nuanced contextual clues related to vocabulary often found in phishing emails. For example, the word "urgent" may be a phishing indicator, but may also be in a legitimate email. The model may not pick up on this difference.

2.4 Suggested Improvements

- Fine-tuning the model on more diverse data may help overcome the limited generalisability problem.

- Making use of a more robust model, like BERT-base, may increase the model's ability to pick up on nuance in its classification. (Although it will require more compute power).

3 Model Results

3.1 Initial Training

These are the results after the fine-tuning script's run had completed.

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1-Score
1	0.036300	0.034065	0.991272	0.996125	0.987086	0.991585
2	0.015500	0.029239	0.993514	0.994523	0.993019	0.993771
3	0.002400	0.035102	0.994666	0.995342	0.994415	0.994878

Two notes on performance:

- Accuracy, Precision, Recall, and F1-scores are above 99%
- Validation loss drops between epoch 1 and 2, which is a good sign of learning, but increases between 2 and 3 which is a sign of minor overfitting.

References

[1] Adrian-Viorel Andriu. "Adaptive phishing detection: Harnessing the power of Artificial Intelligence for enhanced email security". In: *Romanian Cyber Secur. J* 5.1 (2023), pp. 3–9.

[2] Sanjay Ramdas Bauskar et al. "AI-Driven Phishing Email Detection: Leveraging Big Data Analytics for Enhanced Cybersecurity". In: *Library Progress International* 44.3 (2024), pp. 7211–7224.

[3] P. Chinnasamy et al. "AI Enhanced Phishing Detection System". In: *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. 2024, pp. 1–5. DOI: 10.1109/INCOS59338.2024.10527485.

[4] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL]. URL: https://arxiv.org/abs/1910.01108.