COS 720 Project

Linda Masia

April 2025

1 Introduction

Advanced technologies, such as machine learning (ML) and natural language processing (NLP), are used by AI-powered phishing detection systems to more accurately and efficiently detect and stop phishing attempts. [1] ML algorithms analyse vast quantities of data to unearth patterns and anomalies indicative of phishing. NLP techniques enable an understanding of the nuanced language used in phishing communication. Deep learning models, such as CNNs and RNNs, perform complex analyses by automatically extracting features from raw data. Anomaly detection identifies unusual patterns in user behaviour, email traffic, and web interactions that may signal phishing attacks. [2]

AI-based methods have an advantage over traditional rule-based models and blacklists in that they can learn progressively from new data and evolving attack patterns, thereby providing a more dynamic and resilient defence. Traditional methods, based on static rules and known threat lists, cannot address the evolving nature of phishing attacks. AI can identify unknown phishing attacks by identifying subtle patterns and relationships in large datasets.

AI-powered email phishing detection threats involve utilising artificial intelligence (AI) methods to scan email content and metadata to help identify and stop malicious emails that attempt to deceive recipients into disclosing private information. This method is more advanced than keyword matching because it considers the context and purpose of emails.

2 Model Selection Process

The model selection process began by conducting a literature review of the existing models. The BERT model had the best utility. [3] Unlike conventional NLP models, which analyse text one word at a time, the BERT model is built on a transformer architecture that processes the full-text input at once, making it easier to understand the connections between words and sentences. Its high performance in classification tasks and ease of fine-tuning make BERT ideal for phishing detection. [4]

BERT's attention mechanisms make it well-suited for identifying common phrases used in phishing emails. The model can assess the significance of each word in a phrase owing to self-attention, whereas multi-head attention captures different relationships between words for a richer understanding.

After choosing BERT, the training phase began using the Kaggle phishing dataset. We employed checkpointing to save the model state after each epoch, enabling the selection of the best-performing version and early stopping. The phishing dataset was used to fine-tune the classification head and the pre-trained layers. A held-out test set was used for the evaluation. We employed accuracy, precision, recall, and F1-score to measure the performance.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
checkpoint-3133	0.8333	0.8333	0.8333	0.8333
model	0.8333	0.8333	0.8333	0.8333
checkpoint-500	0.5000	0.5000	1.0000	0.6667
checkpoint-1000	0.5000	0.5000	1.0000	0.6667
checkpoint-1500	0.6667	0.6000	1.0000	0.7500
checkpoint-2000	0.7500	0.6667	1.0000	0.8000

Table 1: Performance metrics of phishing detection models

These findings indicate that the model effectively differentiates between phishing and legitimate emails. The refined BERT model achieved reliable detection with few false positives and false negatives, as evidenced by its high accuracy and balanced precision/recall values.

3 Model Evaluation on Email Dataset

Results from running predictions on 1000 email samples on a foreign dataset

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Legitimate	0.92	0.90	0.91	476
Phishing	0.91	0.93	0.92	524
Accuracy		0.92		1000
Macro avg	0.92	0.91	0.91	1000
Weighted avg	0.92	0.92	0.91	1000

Table 2: Classification report on 1000 email samples

3.1 Confusion Matrix

$$\begin{bmatrix} 430 & 46 \\ 39 & 485 \end{bmatrix}$$

4 AI Principles Applied

BERT's ability to pay attention to each word about others was enabled by self-attention and multihead attention. This allowed the detection of subtle patterns in the phishing messages. The model was trained using supervised learning on a labelled dataset of legitimate and phishing emails. We used the accuracy, F1-score, recall, and precision, all of which were obtained from the confusion matrix, to assess the model's efficacy.

5 Limitations and Suggested Improvements

The effectiveness of BERT-based phishing detection depends on the calibration and the variety of training data. The Kaggle dataset may lack examples of complex phishing tactics or emails in multiple languages, limiting its generalisability. Additionally, BERT does not leverage metadata (e.g. sender IP or domain quality), which can enhance accuracy.

To address these limitations, future improvements include the following.

- Augmenting training data with multilingual and diverse phishing examples.
- Incorporating metadata-based features into the model.
- Combining BERT with rule-based systems and anomaly detection to enhance robustness.

6 Diagrams

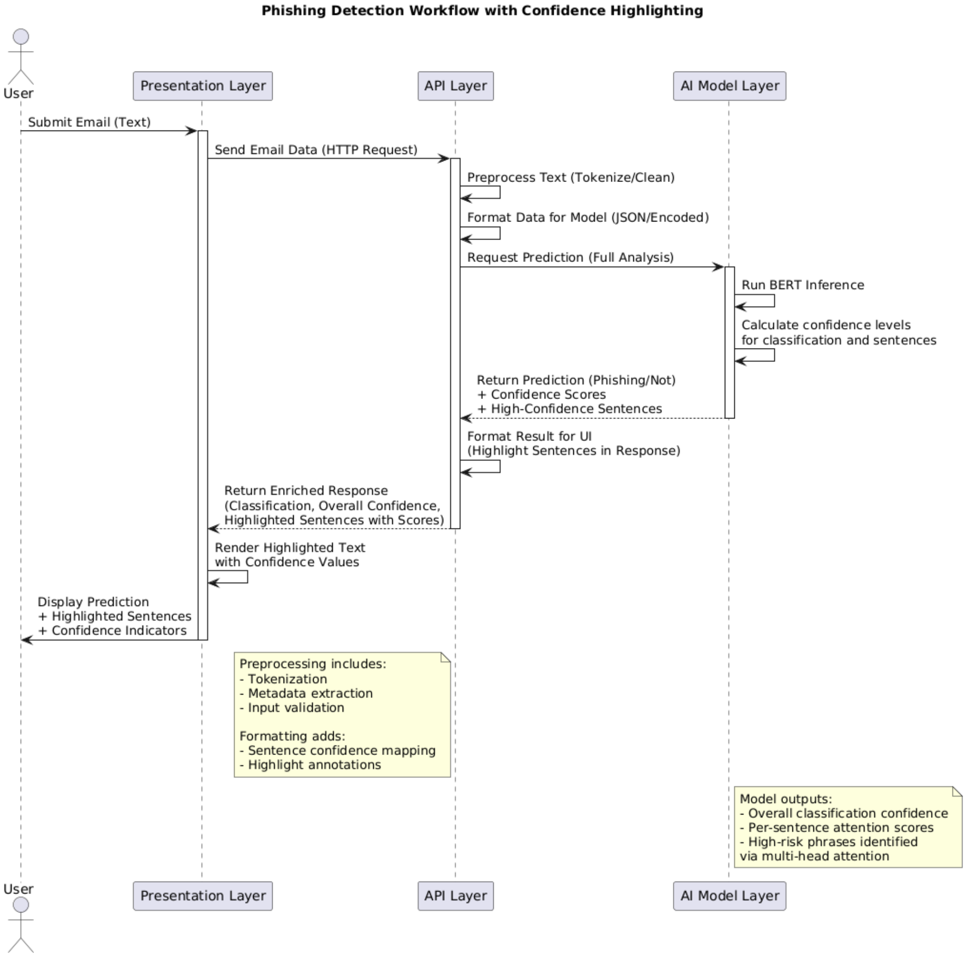
Phishing Detection Workflow with Confidence Highlighting
User
Presentation Layer
API Layer
AI Model Layer
Submit Email (Text)
Send Email Data (HTTP Request)
Preprocess Text (Tokenize/Clean)
Format Data for Model (JSON/Encoded)
Request Prediction (Full Analysis)
Run BERT Inference
Calculate confidence levels for classification and sentences
Return Prediction (Phishing/Not) + Confidence Scores + High-Confidence Sentences
Format Result for UI (Highlight Sentences in Response)
Return Enriched Response (Classification, Overall Confidence, Highlighted Sentences with Scores)
Render Highlighted Text with Confidence Values
Display Prediction + Highlighted Sentences + Confidence Indicators
Preprocessing includes:
- Tokenization
- Metadata extraction
- Input validation
Formatting adds:
- Sentence confidence mapping
- Highlight annotations
Model outputs:
- Overall classification confidence
- Per-sentence attention scores
- High-risk phrases identified via multi-head attention
User
Presentation Layer
API Layer
AI Model Layer

Figure 1: Activity diagram showing the workflow of the phishing detection system

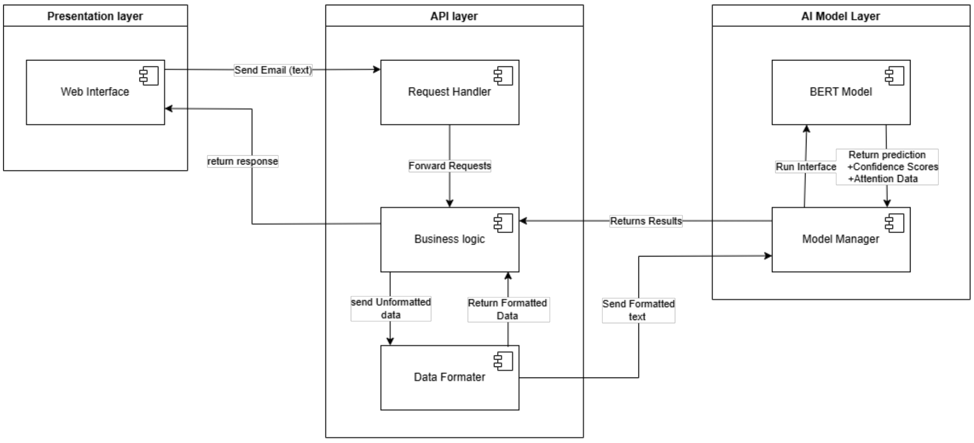

Figure 2: Architecture of the phishing detection system

References

[1] O. I. Enitan, “An ai-powered approach to real-time phishing detection for cybersecurity,” *International Journal*, vol. 12, no. 6, 2023.

[2] B. Gogoi and T. Ahmed, “Phishing and fraudulent email detection through transfer learning using pretrained transformer models,” in *2022 IEEE 19th India Council International Conference (INDICON)*, 2022, pp. 1–6.

[3] M. Songailaitė, E. Kankevičiūtė, B. Zhyhun, and J. Mandravickaitė, “Bert-based models for phishing detection,” in *28th Conference on Information Society and University Studies (IVUS’2023). CEUR Workshop Proceedings. Kaunas, Lithuania*, 2023.

[4] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does bert look at? an analysis of bert’s attention,” *arXiv preprint arXiv:1906.04341*, 2019.