Sniffing for Phishing — UBC Cybersecurity

Proposal for DSCI 591 Capstone Project

Ci Xu Danish Karlin Isa Jia Quan (Joseph) Lim
Lik Hang (Alex) Wong

May 9, 2025

Table of contents

1 Executive Summary

Currently, the University of British Columbia (UBC) Cybersecurity (the Partner) manually reviews suspicious emails reported by users – a process made difficult by the high-volume of reported emails and time sensitivity in addressing these threats.

This project aims to support the Partner in automating the existing review-to-recall workflow. Our solution will enhance threat detection speed, reduce manual workload, and strengthen UBC's cybersecurity.

The deliverables include a containerised web service consisting of a trained machine learning (ML) pipeline that classify reported emails, a dashboard for performance monitoring, and comprehensive documentation.

2 Introduction

Phishing emails represent a significant cybersecurity threat designed to trick recipients into divulging sensitive information including credentials and personal details. According to Internet Crime Complaint Center (2025), business email compromise[1] ranks second in terms of financial damage. At UBC, email users face regular exposure to such phishing attempts, creating substantial risks for both individual users and the university as a whole.

Recipients of such emails can report them to the Partner for further investigation, who reviews and labels them as `CEO Fraud`, `Phishing`, `Spam` or `Legitimate`. `CEO Fraud` and `Phishing` emails are considered malicious and will be recalled from users' mailboxes to prevent further exposure.

2.1 Current works

2.1.1 Classical machine learning models

Classical models demonstrated strong performance for phishing email classification (Meléndez, Ptaszynski, and Masui 2024). These models excel with structured feature sets, learning the importance of different inputs and establishing decision functions for desired classifications (Bergholz et al. 2008). Moreover, classical models offer advantages in interpretability, computational efficiency, and ease of deployment (Meléndez, Ptaszynski, and Masui 2024).

For textual analysis, vectorization techniques such as Bag of Words (BOW) and Term Frequency–Inverse Document Frequency (TF-IDF) are commonly used. These methods are compatible with classical models by converting texts into numerical representations (Meléndez, Ptaszynski, and Masui 2024).

2.1.2 Deep learning models

Multiple works leveraged transformer-based architectures like Bidirectional Encoder Representations from Transformers (BERT) to obtain rich, contextual representations of the subject and body using attention-based bidirectional learning (Otieno et al. 2023; Otieno, Siami Namin, and Jones 2023). Deep learning models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks and transformer-based models have also shown strong performance in detecting phishing emails (Shazad et al. 2024; Otieno, Siami Namin, and Jones 2023; Bagui et al. 2019).

[1] Business email compromise is a form of scam that targets users in an organisation and seek information with the intention of defrauding the organisation.

2.1.3 Large language models

There is existing work that leveraged large language models (LLMs) for email classification through prompt engineering. Chain-of-thought prompting was used where the complex task of email classification is decomposed into smaller, manageable problems (Koide et al. 2024). This approach has shown to improve the quality of the LLM's output by encouraging the model to reason (Wei et al. 2022). Additionally, this method managed to outperform classical and deep-learning methods.

2.2 Current state

Figure 1 describes the current review-to-recall process.

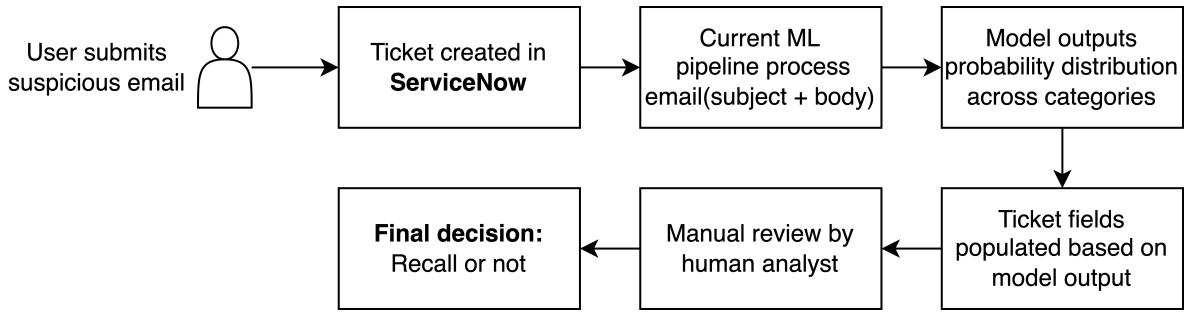

Figure 1: Current review-to-recall process

Currently, the performance of the ML pipeline that the Partner uses is not sufficient to support automation of the review-to-recall process. Figure 2 shows the current ML pipeline.

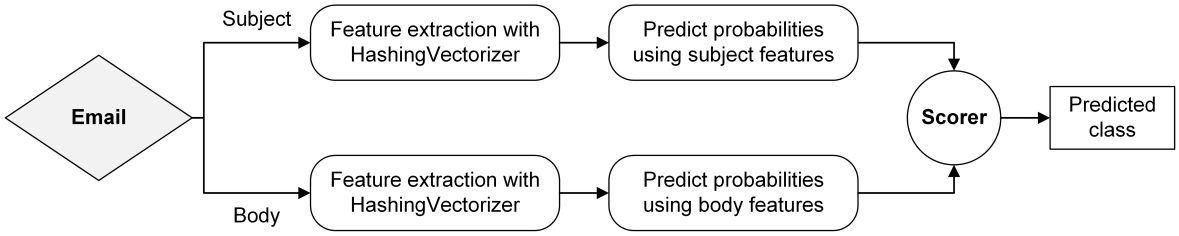

Figure 2: Current ML pipeline used by the Partner

2.3 Objectives

We aim to reduce the Partner's response time to malicious email campaigns with a new ML pipeline that supports full automation of the review-to-recall process through:

- expanding feature extraction to better use of the email metadata and content, and
- leveraging classical, deep learning techniques, and LLMs to classify emails.

2.4 Deliverables

The deliverables for this project are:

1. A ML pipeline that parses emails and generates a predicted class
2. A containerised web service hosting the pipeline and callable via a RESTful API
3. A dashboard for performance monitoring
4. Comprehensive documentation

The minimum viable products for this project are deliverable nos. 1 and 4.

3 Data Science Techniques

Figure 3 outlines the workflow for this project.

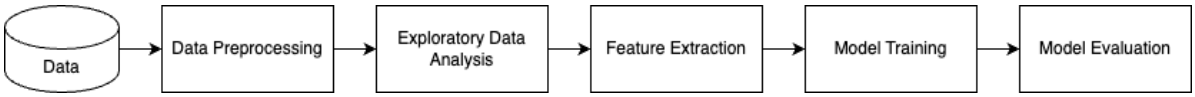

Figure 3: Project workflow

3.1 Dataset and data structure

The Partner provided a collection of reported emails which come in various formats. However, we will only work with `eml` files. The file structure of the dataset is shown below.

```
data
   label
       ticket_id
           file-1
           ...
           file-n
```

Figure 4 shows the file structure of an `eml` file.

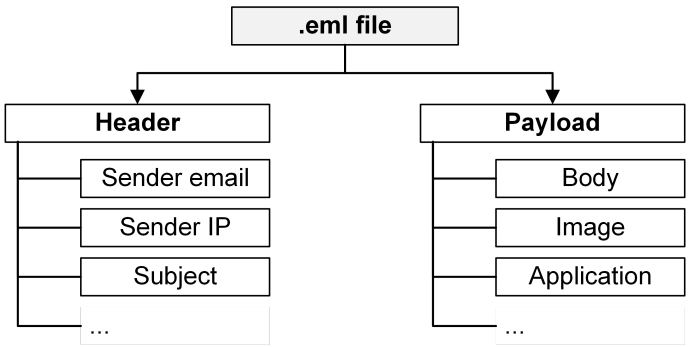

Figure 4: Structure of an `eml` file

We will parse the emails using the `email` package and extract metadata from the header, such as the sender, subject, and body. Additionally, we will use tools like `BeautifulSoup` to extract text from rich-text email bodies in `html` format.

3.2 Modelling approaches

We propose three different approaches to meet our objectives.

3.2.1 Classical machine learning

We first leverage a range of classical models, including Naïve Bayes, Logistic Regression (LR), and Support Vector Machine (SVM), for our classification task.

For better performance, we aim to extract a comprehensive set of features from the emails, including metadata such as sender information and routing details. Additionally, we will analyze email payload such as text, presence of links, and various characteristics of embedded links. We will process text with BOW and TF-IDF and extract latent features using techniques such as Latent Dirichlet Allocation (LDA).

Figure 5 shows the pipeline for this approach.

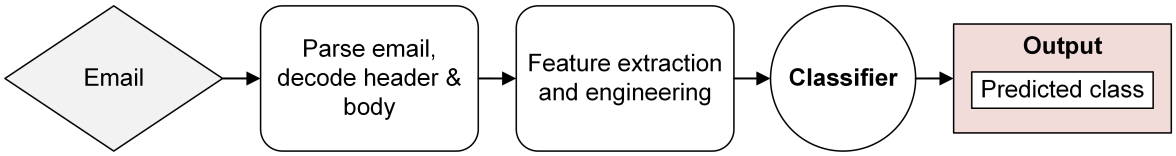

Figure 5: Pipeline for classical and deep learning approaches

3.2.2 Deep learning

Our second approach leverages transformer-based architectures, specifically BERT, to extract features from text and generate meaningful contextual embeddings. These embeddings will serve as input to deep learning models such as CNN and LSTM, which are well-suited for the classification task due to their ability to learn complex patterns and identify anomalies within the data. Additionally, we will also experiment with transformer-based models.

The pipeline for this approach similar to the classical approach shown in Figure 5.

3.2.3 Large language models

We intend to explore the use of LLMs for classifying emails using chain-of-thought prompting. This involves generating a structured query that encourages the LLM to justify its classification. Unlike other approaches, LLMs do not require extensive feature extraction and may be more sensitive to anomalies in the emails.

Figure 6 shows the pipeline for this approach.

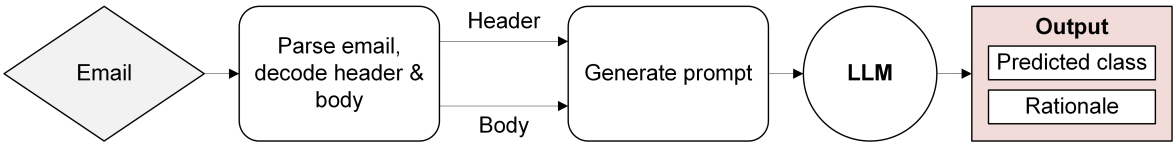

Figure 6: Pipeline for the LLM approach

Due to data residency constraints, we intend to work with open-source LLMs that can be downloaded and trained locally on the provided virtual machine (VM).

3.3 Evaluation of proposed pipelines

The current ML pipeline shall serve as a baseline for evaluation. The evaluation metrics are F1-score and False Positive Rate (FPR)[2], as specified by the Partner.

Specifically, the proposed pipeline shall achieve F1-score ≥ 0.85 and FPR ≤ 0.001 to enable the automated recall of suspected malicious emails. A high F1-score indicates the model correctly identifies most phishing emails while minimizing false positives, while maintaining a low FPR is crucial to avoid recalling legitimate emails by mistake, preventing important messages from being missed.

4 Conclusion

In this project, we aim to tackle the threat of malicious emails at UBC by developing a high-performing ML pipeline capable of classifying reported emails as malicious or legitimate.

For the Partner, this project supports efforts to automate the review-to-recall process for reported emails. Currently, analysts manually review reported emails and decide whether to recall similar emails from other mailboxes. This process is made worse with the high volume of incoming tickets. As such, a pipeline with excellent performance can review and automatically recall such emails. This would shorten the response time to malicious email campaigns and free up time for analysts to attend to other tasks.

UBC Email users will also benefit as they are the primary targets of malicious email campaigns. As recipients of such emails, they are at risk of compromising sensitive information or suffering financial loss. A high-performing model would enable the Partner to respond to these threats quickly, reducing the number of UBC users exposed to such campaigns.

[2]Refer to Appendix for definition.

5 Timeline

Table 1 describes the proposed timeline for this project.

Table 1: Proposed timeline with milestones

Milestone	Due	What we'll work on
1	May 9	Final proposal
2	May 16	**Classical approach:** Feature extraction and engineering, training and tuning, evaluation
3	May 23	**Deep learning approach:** Feature extraction and engineering, training and tuning, evaluation
4	May 30	**LLM approach:** Feature extraction and prompt engineering, training and tuning, evaluation
5	June 6	Code refactoring and training on full dataset, packaging our pipeline in a container accessible via an API, development of dashboard for performance monitoring
6	June 12	Final presentation
7	June 18	Draft final report
8	June 25	Final report, packaged data product

6 Appendix

F1-score is the harmonic mean of precision (fraction of true positives out of all positive predictions) and recall (fraction of true positives out of all positives). It combines both into a single score that balances the trade-off between false positives and false negatives.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where,

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

False positive rate measures the fraction of false positives out of all negative examples.

$$\text{False Positive Rate (FPR)} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

References

Bagui, Sikha, Debarghya Nandi, Subhash Bagui, and Robert Jamie White. 2019. "Classifying Phishing Email Using Machine Learning and Deep Learning." In *2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, 1–2. Oxford, United Kingdom: IEEE. https://doi.org/10.1109/CyberSecPODS.2019.8885143.

Bergholz, Andre, Gerhard Paaß, Frank Reichartz, Siehyun Strobel, and Jeong-Ho Chang. 2008. "Improved Phishing Detection Using Model-Based Features."

Internet Crime Complaint Center. 2025. "Internet Crime Report 2024." https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf.

Koide, Takashi, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. 2024. "ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection." arXiv. https://doi.org/10.48550/arXiv.2402.18093.

Meléndez, René, Michal Ptaszynski, and Fumito Masui. 2024. "Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection." *Electronics* 13 (24): 4877. https://doi.org/10.3390/electronics13244877.

Otieno, Denish Omondi, Faranak Abri, Akbar Siami Namin, and Keith S. Jones. 2023. "Detecting Phishing URLs Using the BERT Transformer Model." In *2023 IEEE International Conference on Big Data (BigData)*, 2483–92. Sorrento, Italy: IEEE. https://doi.org/10.1109/BigData59044.2023.10386782.

Otieno, Denish Omondi, Akbar Siami Namin, and Keith S. Jones. 2023. "The Application of the BERT Transformer Model for Phishing Email Classification." In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1303–10. Torino, Italy: IEEE. https://doi.org/10.1109/COMPSAC57700.2023.00198.

Shazad, Ashiq, Muhammad Naman Chaudhry, Muhammad Kamran Abid, and Naeem Aslam. 2024. "Spam Email Detection Using Transfer Learning of BERT Model."

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models."