Methods

Research Question

What characteristics make an email most likely to be a phishing attempt?

Data and Observations

We use the TREC_06 dataset from Kaggle, which consists of approximately 15,000 emails, each labeled as either phishing (1) or not (0). In this study, each observation is a single email, represented as a row containing features such as body text, subject line, sender, presence of URLs, and timestamp. Since our target variable is binary, this study will use supervised learning for classification rather than regression.

Models and Algorithms

To classify phishing emails, we will experiment with models including Logistic Regression, Random Forest Classifier, and SVM with TF-IDF Vectorization--each of which serves a different purpose. Logistic Regression is a simple classification approach that will estable a baseline for prediction performance. Random Forest Classifier can handle structured features like presence of URLs, email length, and sender frequency while simultaneously reducing overfitting. SVM with TF-IDF Vectorization allows us to extract textural patterns from email bodies using Term Frequency-Inverse Document Frequency (TF-IDF) to convert text into numerical features, like the email body, subject line, and URLs, and then use SVM to analyze those textural features. This aids in capturing the importance of specific words within the context of the email while reducing the impact of more common words including "and", "in", and "the".

Feature Engineering

Given that phishing emails rely heavily on textual deception, Natural Language Processing (NLP) techniques will be employed to refine the dataset. Specifically, we will experiment with different NLP techniques such as stemming, lemmatization, and n-grams. Furthermore, the data may require additional cleaning to fix inconsistent timestamps, missing values, or encoding errors in emails.

Model Evaluation and Success Metrics

Success is defined by a model that achieves high predictive accuracy without significant overfitting. Success of the model was measured through accuracy (overall classification correctness), precision and recall (missed phishing email classifications), F1 score (balance between precision and recall), and ROC-AUC score (measure of overall classifier performance). Precision and recall is particularly important for phishing detection since false negatives are riskier than false positives, as these emails can contain potentially dangerous links that steal pertinent receiver information. Visualizations such as confusion matrices and feature importance plots will provide deeper insights into model performance.

Anticipated Challenges & Mitigation Strategies

Feature engineering may require multiple iterations. We will experiment with different NLP techniques to improve classification accuracy. Additionally, inconsistent timestamps, missing values, and encoding errors may arise. We will preprocess data to address these issues before model training. Finally, regularization techniques and cross-validation will be used to ensure models generalize well to unseen data. If our approach does not achieve the desired

performance, we may gain insights into identifying additional phishing indicators,
May find ourselves exploring deep learning models like LSTMs or BERT, or understanding evolving tactics used in phishing campaigns. By iterating on these approaches, we aim to develop an interpretable model for phishing email detection.

Results

Once the models are trained and evaluated, we will present our findings through various visualizations and performance metrics. We will use confusion matrices to analyze misclassification patterns and evaluate false positives and false negatives, which are critical in phishing detection. Similarly, we will have a summary table comparing accuracy, precision, recall, F1-score, and ROC-AUC for each model to determine the best-performing classifier. Additionally, we will look into the importance of given features for the best performing classification algorithm. Through these results, we aim to answer which characteristics most strongly indicate phishing emails and determine which model is best suited for practical phishing detection applications.