# Detecting Phishing Emails

Lincoln Brown

DSC680-T301

Professor Iranitalab

## Topic

Detecting Phishing Emails. This project will develop a machine learning model to classify emails as safe or phishing based on the text in the email. The goal is to create a model that can be used to help detect phishing emails to improve email security using Natural Language Processing (NLP).

## Business Problem

Emails are a popular platform for attackers to gain access to a network. Using phishing emails, malicious actors harvest user credentials which they can then use to gain access to sensitive information or to impersonate the end user and continue their attack on associated businesses and entities. Phishing emails are a wide-spread business problem and can be a critical vector of attack for data breaches. This project is focused on producing a model that can be used to strengthen phishing detection using models trained on data obtained from Kaggle.

## Datasets

The dataset used in this project is the Phishing Email Detection found on Kaggle, uploaded by user Cyber Cop. This dataset contains two features: Email Text and Email Type. The email text feature contains the body of the email, and the email type is the associated classification label phishing or safe.

## Methods

This project will build two models which will then be compared for performance. The first will be a Logistic Regression model built using Term Frequency-Inverse Document Frequence (TF-IDF). The second model will be a Bidirectional Encoder Representations from Transformers (BERT) model.

Additionally, the models will be evaluated using accuracy, precision, recall, F1-score, Area Under the Curve – Receiver Operating Characteristic (AUC-ROC), and a confusion matrix to determine which model is the most effective.

If problems are encountered with building the BERT model, it may be substituted for a different model such as a Long-Short Term Memory (LSTM) model or a Naïve Bayes model.

## Ethical Considerations

The ethical considerations for this project are fairly minimal. The dataset is public and does not contain personally identifiable information (PII) such as sender and receiver email addresses. Furthermore, the dataset is limited to just two features, the email body and the email type, which keeps the data strictly focused on classification.

## Challenges/Issues

The dataset is slightly imbalanced, with 61% being labelled as safe emails and 39% being labelled as phishing emails. This imbalance could lead to a bias towards the safe class. This will be monitored, and corrective steps will be taken if necessary.

Additionally, a key challenge in this project is the lack of experience with building and tuning a BERT model, which may take additional time for experimentation and learning. If the BERT model is unsuccessful, a Long-Short Term Memory model or Naïve Bayes model may be used instead.

## References

A textbook that contains guides for employing TF-IDF.

Albon, C. (2018). *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*.

The dataset from Kaggle.

Chakraborty, A. N. (n.d.). *Phishing email Detection*. https://www.kaggle.com/datasets/subhajournal/phishingemails?select=Phishing_Email.csv

A guide to implementing text classification using BERT.

Erim, E. (2024, March 26). Complete guide to building a text classification model using BERT. *Medium*. https://medium.com/@emreerim_65318/complete-guide-to-building-a-text-classification-model-using-bert-abf27b5cb6a1

An article that describes Python's NLTK package for NLP.

Jablonski, J. (2023, October 21). *Natural language processing with Python's NLTK package*. https://realpython.com/nltk-nlp-python/