CPS579: Final Project

Detecting Phishing/Spam using Transformer Models
Total Marks: 30 Points

Due Date: April 30, 2025

Overview

In this final project, you will leverage **Hugging Face's Transformer models** to design and implement a machine learning pipeline that can detect malicious content in the form of:

- **Phishing Emails**
- **Spam SMS Messages**
- **Phishing URLs**

You will apply Natural Language Processing (NLP) techniques using pre-trained models such as `BERT`, `RoBERTa`, or `DistilBERT`, and fine-tune them on relevant datasets. This hands-on project will help you understand how advanced language models can be used for real-world cybersecurity applications.

Objectives

- Explore and preprocess datasets containing phishing or spam content
- Fine-tune two Hugging Face transformer models for binary classification (malicious vs. legitimate)
- Evaluate model performance using standard metrics (accuracy, precision, recall, F1-score)
- Plot confusion matrices for each model validation
- Interpret model behavior and discuss real-world applications in security
- plot a bar plot showing accuracy comparison between two models

Tasks

1. Choose one detection task: **Phishing Email**, **Spam SMS**, or **Phishing URL**
2. Collect or use an open-source dataset (e.g., SMS Spam Collection, Phishing Email datasets from Kaggle, PhishTank URLs)
3. Preprocess the text data: cleaning, tokenization, and label encoding
4. Fine-tune a Hugging Face transformer model using the `transformers` and `datasets` libraries
5. Evaluate each model using a held-out test set and optionally perform cross-validation
6. Interpret results using attention visualizations. Plot confusion matrices for each model validation and plot a bar plot showing accuracy comparison between two models

Deliverables

- A Jupyter Notebook or Python script with a complete ML pipeline
- A short report or presentation that includes:
  - Problem statement
  - Dataset summary
  - Model architecture and training setup
  - Evaluation metrics and discussion
  - Challenges and possible improvements

Tools and Resources

- Python, Jupyter Notebook
- `transformers`, `datasets`, `scikit-learn`, `pandas`, `matplotlib`
- Hugging Face Model Hub: `https://huggingface.co/models`
- Hugging Face Text model example: `https://huggingface.co/docs/transformers/en/model_doc/bert#transformers.TFBertTokenizer`