Week 5 Direct Independent Study Progress

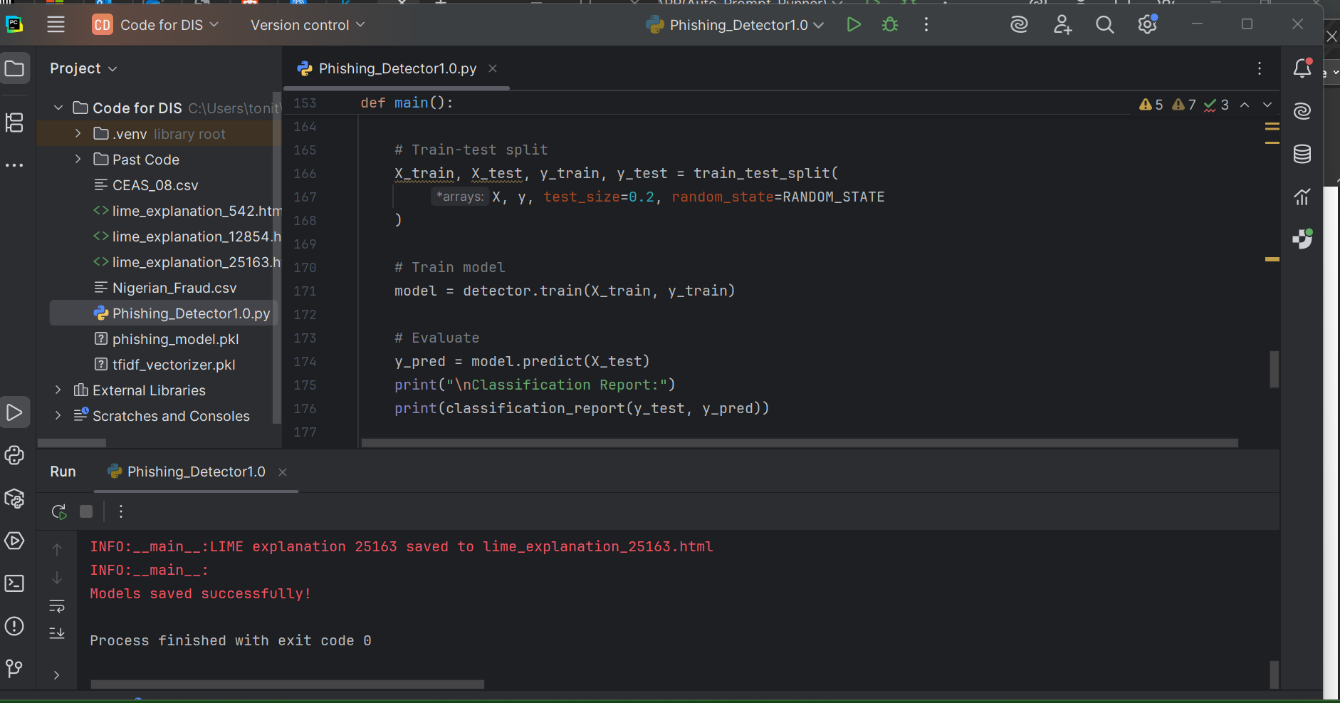
* **Topic Covering**: Al-Powered Phishing Detection Using NLP & Explainable Al.
  + **Covered Paper: Deep Learning for Phishing Detection: Taxonomy,**
  + **Current Challenges and Future Directions**
    - **Link**: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9716113>
    - **Citation:**  *Alghowinem, Sharath, Nour Moustafa, Benjamin Turnbull, and Ernest Foo. 2021. "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions." Computers & Security 105: 102992.* <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9716113>
    - **Notes**: In this paper, the authors compare traditional ML (Random Forest, SVM) vs deep learning (LSTM, BERT, Transformers), highlighting NLP-based approaches (TF-IDF, word embeddings) for email/text analysis.
      * They also discuss SHAP/LIME for interpreting model decisions and find that SHAP is more stable for global explanations, while LIME excels in local interpretability.
      * They evaluate performance on legitimate vs phishing datasets and emphasize class imbalance as a major challenge.
      * A hybrid model is one of the recommendations the authors make for better context awareness.

**Relevance of Paper to my Study**

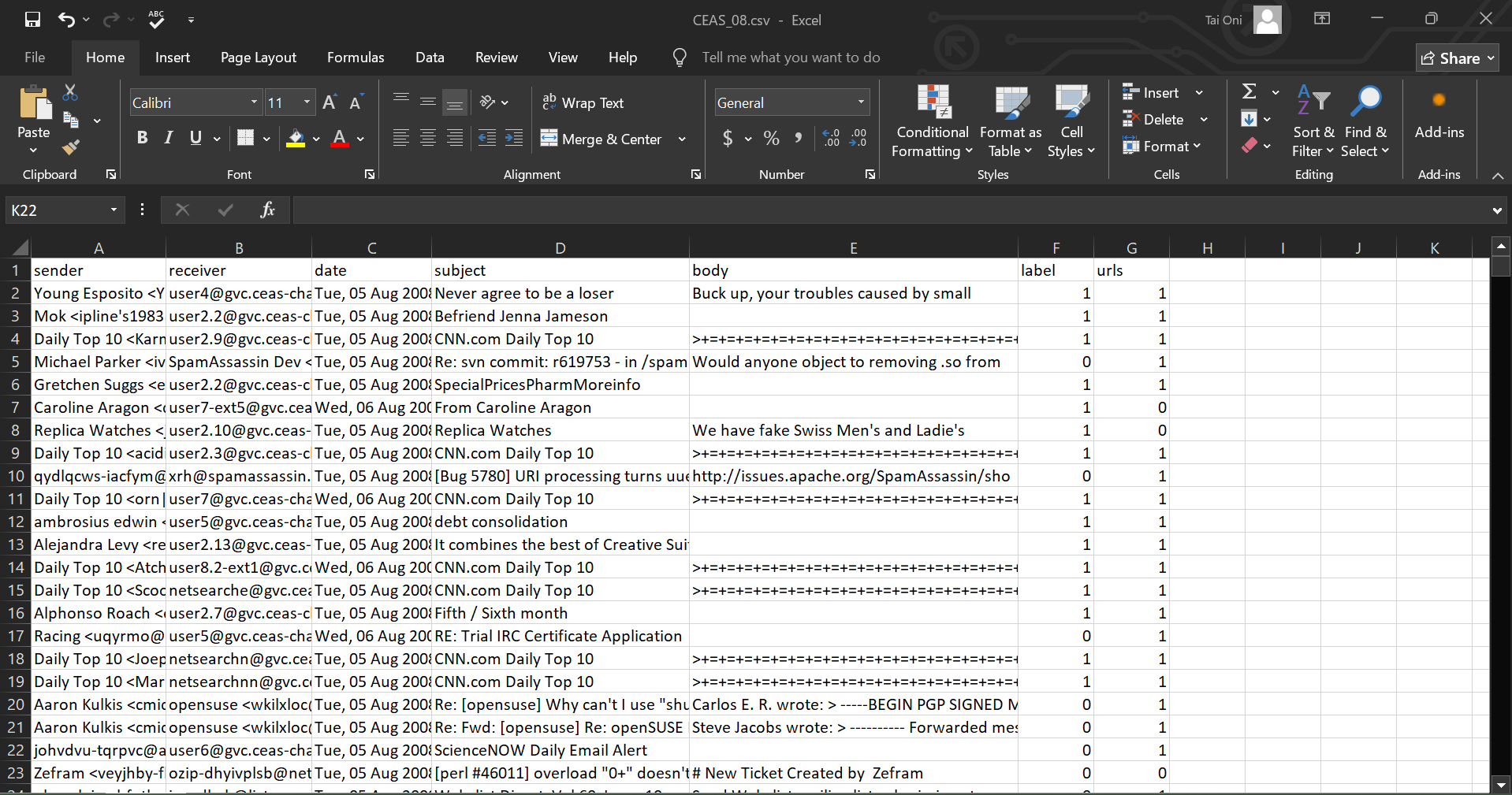
* + The first paper validates your TF-IDF + Random Forest pipeline as well as supports my use of SHAP/LIME for explainability.

# **Data Preprocessing**

* ***Code from detecting phishing email has commenced.***

******

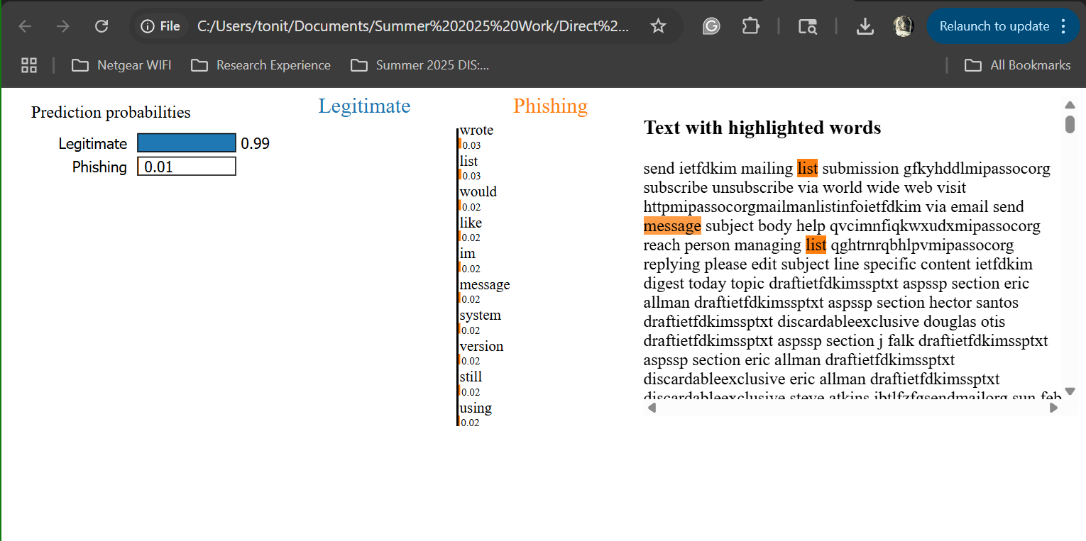
* ***CEAS-08 Fraud File csv data used for initial test run***

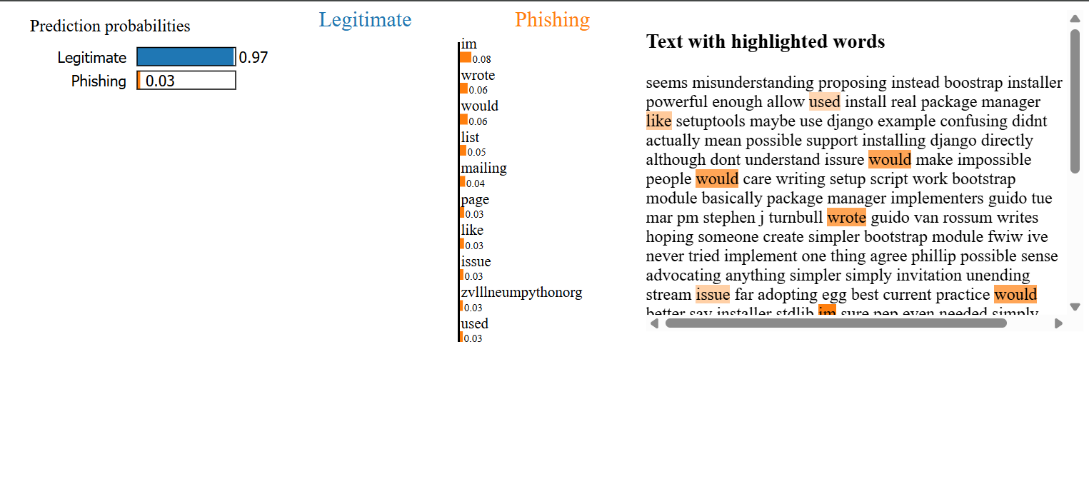
******

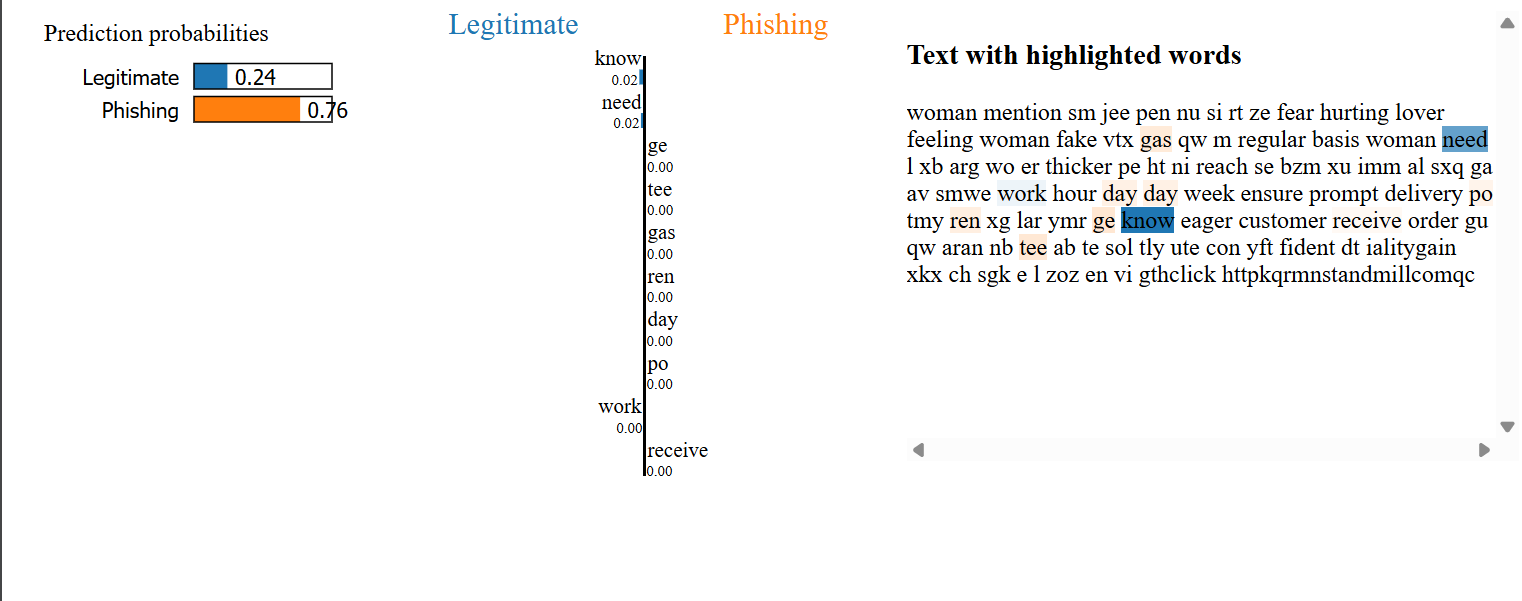
* ***Results:*** 
  + ***Code successfully parses the csv data and used the information to train the model***
    - ***Loaded 39154 emails (21,842 were Phishing & 17,312 were legitimate)***
    - ***Features:***

|  |  |
| --- | --- |
| ***Sender*** | ***Email Sender*** |
| ***Receiver*** | ***Email Recipient*** |
| ***Date*** | ***TimeStamp*** |
| ***Subject*** | ***Email Subject Line*** |
| ***Body*** | ***Email Content*** |
| ***url*** | ***Embedded URLs*** |
| ***Label*** | ***1 = Phishing, 0 = Legitimate*** |

* + - ***Preprocessing used:***
      * ***Removed missing values (body field required).***
      * ***Text cleaning (lowercase, special character removal).***
      * ***Tokenization, stopword removal, lemmatization.***
  + ***Proposed Methodology:*** 
    - ***Machine Learning Pipeline: Feature Extraction 🡪 Model Training 🡪 Unsupervised Learning***
    - ***Explainable AI (XAI): SHAP (Shapley Additive Explanations) & LIME (Local Interpretable Model-Agnostic Explanations 🡪 Visualizations (Next step)***
      * ***Converted email text into numerical features.***
      * ***Used Random Forest Classifier (100 trees, max depth=10)***
      * ***SHAP Identifies key words influencing predictions.***
      * ***It is optimized for performance (batched processing, sample size control).***
      * ***Code is designed to fallback to KernelSHAP if TreeSHAP fails.***
      * ***LIME generates per-email explanations (HTML output).***
      * ***Highlights top 10 influential words per prediction.***

******

******

******

* ***Still working on code as SHAP generation runs into errors. Plots are also not being generated.***