Week 3 Direct Independent Study Progress

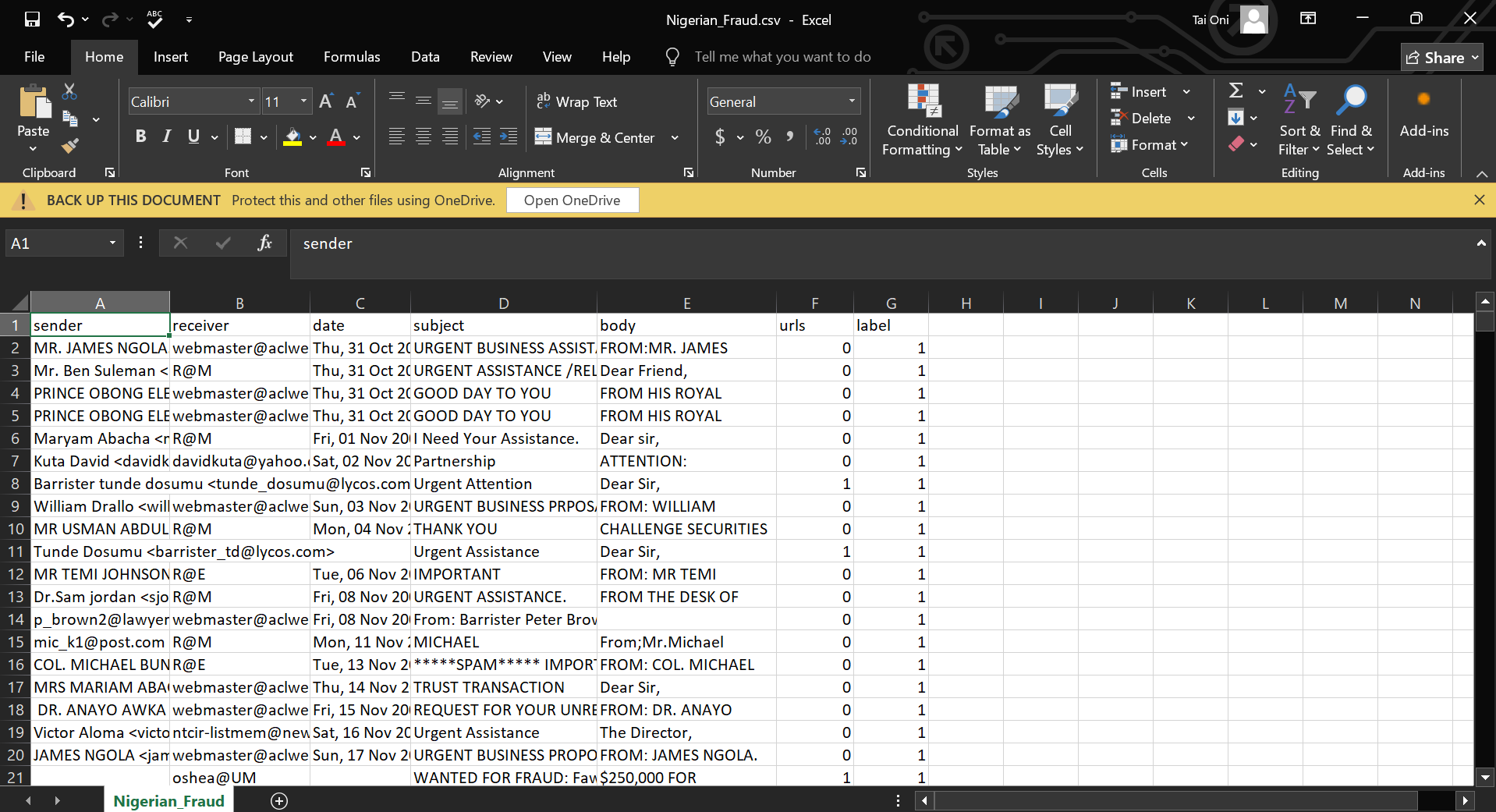
* **Topic Covering**: Al-Powered Phishing Detection Using NLP & Explainable Al.
* **Covered Paper:** In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets
  + **Link**: <https://www.researchgate.net/publication/390030930_In-Depth_Analysis_of_Phishing_Email_Detection_Evaluating_the_Performance_of_Machine_Learning_and_Deep_Learning_Models_Across_Multiple_Datasets>
  + **Citation:**
  + **Notes**: This study involves a fully detailed evaluation of different machine learning and deep learning models for phishing email detection. Using a handpicked dataset that includes emails from SpamAssassin, UCI Machine Learning Repository, among others, the authors of this paper evaluate models such as CNN, XGBoost, RNN, SVM, and a hybrid BERT-LSTM model.
    - The BERT-LSTM model achieved the highest performance, with an accuracy of 99.55% and an F1-score of 99.24%. The study highlights the superiority of transformer-based models in capturing complex patterns in phishing emails.
* **Covered Paper:** Improving Phishing Email Detection Performance of Small Large Language Models
  + **Link**: <https://arxiv.org/abs/2505.00034>
  + **Citation:**
  + **Notes**: The authors of this paper investigate ways to enhance the phishing email detection capabilities of small large language models (LLMs) with approximately 3 billion parameters.
    - They address the challenges of deploying resource-intensive models by offering techniques such as prompt engineering, explanation-augmented fine-tuning, and model ensembling.
    - Their experiments which were done on datasets like CEAS\_08 and SpamAssassin, showed that these approaches significantly improve the performance of small LLMs, making them a viable option for phishing detection tasks without the need for extensive computational resources.

**Relevance of Paper to my Study**

* The first paper showed the effectiveness of transformer-based models like BERT-LSTM in phishing detection while also giving a benchmark for model performance using datasets similar to what I am using.
* The second paper gave strategies to optimize smaller LLMs for phishing detection, aligning with practical deployment considerations. It also brings up explanation-augmented fine-tuning, which can potentially enhance my model's interpretability.

# **Data Preprocessing**

* Dataset used was found from Kaggle ([Phishing Email Dataset](https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset)). It contained a combination of emails from a variety of sources to create a comprehensive set with the following:
  + Initial Dataset:
    - **Enron and Ling Datasets**: These datasets focus on the core content of phishing emails, containing subject lines, email body text, and labels indicating whether the email is spam (phishing) or legitimate.
    - **CEAS, Nazario, Nigerian Fraud, and SpamAssassin Datasets:** These datasets provide broader context for the emails, including sender information, recipient information, date, and labels for spam/legitimate classification.



* + Final Dataset: phishing\_email.csv combines the information from the initial datasets into a single resource for analysis. This dataset contains:
    - Approximately 82,500 emails
    - 42,891 spam emails
    - 39,595 legitimate emails
    - The "text\_combined" column is the central element of the final dataset in this phishing email analysis. It combines the subject line, the body, date, and sender email text of the emails from the initial datasets.
* **NOTE: Taiwo, remember to cite the following article if using this dataset**
  + ***\*Al-Subaiey, A., Al-Thani, M., Alam, N. A., Antora, K. F., Khandakar, A., & Zaman, S. A. U. (2024, May 19). Novel Interpretable and Robust Web-based AI Platform for Phishing Email Detection. ArXiv.org. https://arxiv.org/abs/2405.11619\****