Week 1 Direct Independent Study Progress

* **Topic Covering**: Al-Powered Phishing Detection Using NLP & Explainable Al.
* **Covered Paper: An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach**
  + **Link**: [[2402.13871] An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach](https://arxiv.org/abs/2402.13871)
  + **Citation:** *"An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach."*  
    arXiv preprint arXiv:2402.13871, February 2024.  
    <https://arxiv.org/abs/2402.13871>.
  + **Notes**: The author’s goals are to develop a lightweight, transformer-based model for phishing email detection, incorporating explainable AI techniques to interpret model predictions. Their method of approach taken involves:
    - Data Preprocessing:
      * Removed null values from the dataset.
      * Addressed class imbalance by oversampling the minority class (phishing emails) to achieve a balanced dataset.
    - Model Selection and Optimization:
      * Utilized DistilBERT, a smaller and faster variant of BERT, as the base model.
      * Conducted hyperparameter tuning:
        + Batch sizes of 8, 16, and 32 were tested; 32 were selected for training.
        + Learning rates of 1e-5, 2e-5, and 3e-5 were evaluated; 2e-5 were chosen.
        + Trained the model for 6 epochs using the AdamW optimizer.
    - Explainability Analysis:
      * + Applied LIME (Local Interpretable Model-Agnostic Explanations) to provide local explanations for individual predictions by perturbing input data and observing changes in model output.
        + Used Transformer Interpret to analyze attention weights, identifying which parts of the input the model focused on during prediction.

The authors based on their test concluded that the fine-tuned DistilBERT model achieved high accuracy in classifying phishing emails while explainability techniques provided insights into model decision-making, highlighting influential words or phrases in emails.

* **Covered Paper**: **EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability**
  + **Link**: [[2503.20796] EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability](https://arxiv.org/abs/2503.20796)
  + **Citation**: *"EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability."*  
    arXiv preprint arXiv:2503.20796, March 2025.  
    <https://arxiv.org/abs/2503.20796>.
  + **Notes**: The authors introduce EXPLICATE, a framework that combines machine learning-based phishing detection with explainable AI techniques and large language models to generate user-friendly explanations of model decisions. Their methodology involves implementing a machine learning classifier using domain-specific features for phishing detection with integrated dual-explanation layer combining LIME and SHAP to provide complementary insights into model predictions. They also employed DeepSeek v3, a large language model, to translate technical explanations into accessible natural language narratives.
    - They also developed a fully usable GUI application and a lightweight Chrome extension to demonstrate the framework's applicability in various deployment scenarios.
    - Results drawn showed 98.4% accuracy in phishing detection, comparable to existing deep learning techniques.
    - Generated high-quality explanations with 94.2% accuracy and 96.8% consistency between the LLM output and model predictions.

The author’s research enhances transparency in phishing detection systems, bridging the gap between automated AI and user trust. It also provides accessible explanations, aiding users in understanding and responding to phishing threats.