PHÁT HIỆN EMAIL LỪA ĐẢO DẪN ĐẾN RANSOMWARE BẰNG MÔ HÌNH TRANSFORMER NHẸ

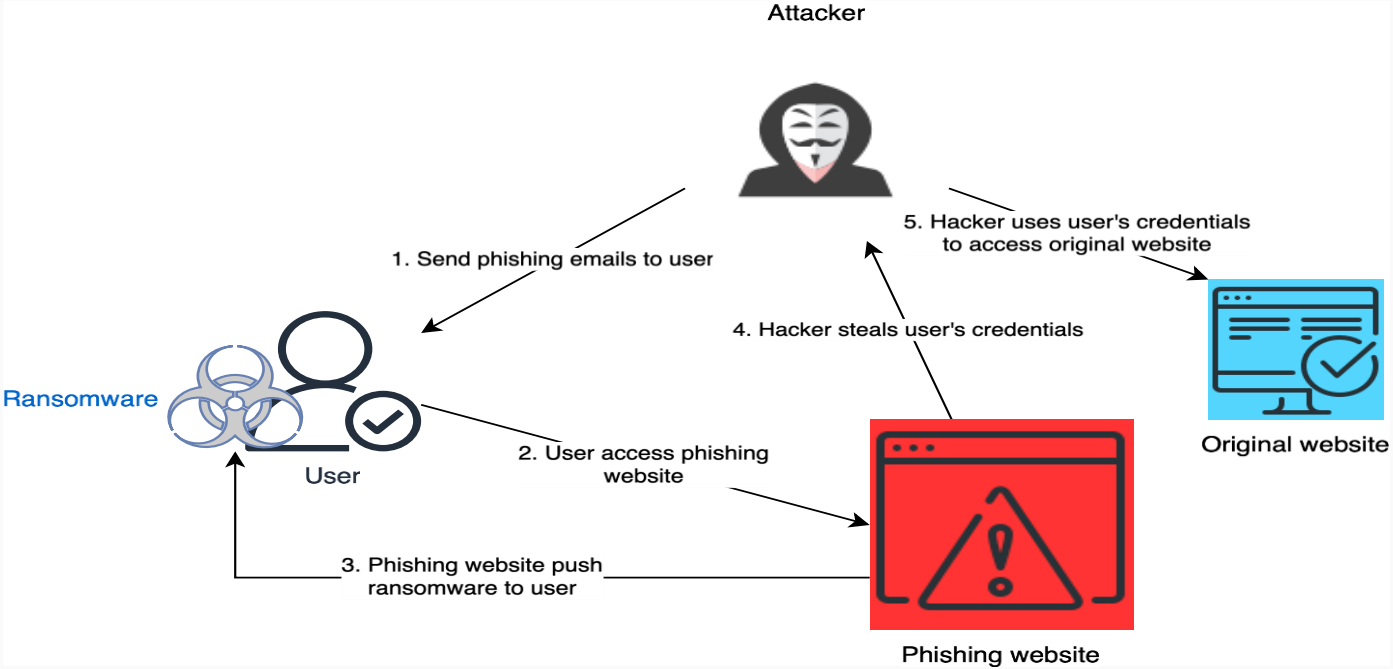
Tạ Duy Huy - 240202023

Tóm tắt

- Lớp: CS2205.FEB2025
- Link Github của nhóm: https://github.com/tdHuy22/CS2205.FEB2025.git
- Link YouTube video: https://youtu.be/a0hVoal02dE

Giới thiệu

- **Bối cảnh:** Email lừa đảo là cửa ngõ chính đến lừa đảo, ransomware tống tiền gây thiệt hại 18,7 tỷ USD năm 2023.



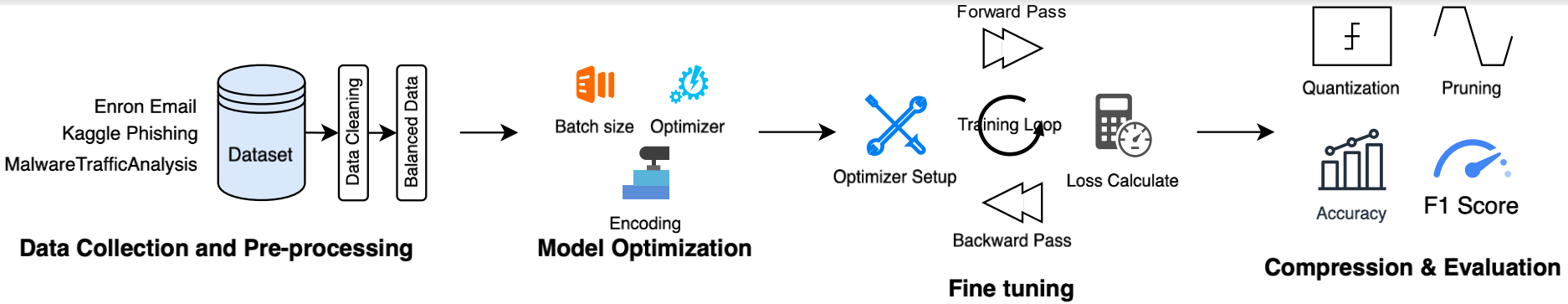
Giới thiệu

- **Hạn chế:**
  - SVM, Random Forest: Khó phân tích ngữ cảnh phức tạp.
  - BERT: Yêu cầu tài nguyên lớn.
- **Đề xuất:** DistillBERT nhẹ, tối ưu cho laptop/Raspberry Pi, tích hợp LIME.

Mục tiêu

- Xây dựng mô hình phân loại email lừa đảo với F1-score > 95%.
- Tối ưu hoá mô hình (< 100MB, suy luận < 1 giây) cho laptop và Raspberry Pi.
- Cung cấp giải thích dự đoán bằng LIME, đạt điểm dễ hiểu > 4/5.

Nội dung và Phương pháp



Thu thập và tiền xử lý dữ liệu:

- Nguồn: Enron Email Dataset, Kaggle Phishing Dataset, MalwareTrafficAnalysis.
- Tiền xử lý: Loại bỏ HTML bằng BeautifulSoup, mã hoá bằng BERT Tokenizer, xử lý mất cân bằng lớp bằng class weighting.

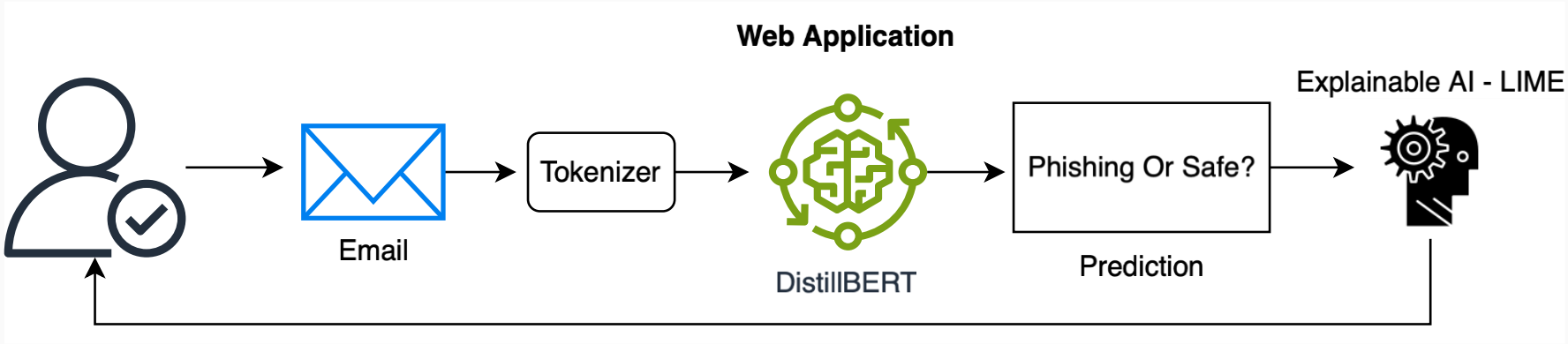
Huấn luyện (Fine-tuning): Fine-tune DistillBERT trên Colab (GPU Tesla T4), batch size: 16, learning rate: 2e-5, 5 epoch

Đánh giá và tối ưu hoá mô hình:

- Đánh giá: F1-score, FPR
- Quantization, pruning, ONNX Runtime

Nội dung và Phương pháp

- **Tích hợp khả năng giải thích XAI:**
  - LIME xác định từ khoá quan trọng (e.g., “urgent”: 0.75)
  - Đánh giá độ dễ hiểu qua khảo sát, mục tiêu > 4/5
- **Triển khai:** Triển khai trên Flask, thử nghiệm trên laptop (RAM 8GB) và Raspberry Pi



Kết quả dự kiến

- **Hiệu suất:** F1-score > 95%, FPR < 1%.
- **Triển khai:**
  - Mô hình < 100MB, suy luận < 1 giây.
  - Hoạt động trên laptop, Raspberry Pi.
- **Minh bạch:** LIME đạt điểm dễ hiểu 4/5.
- **Ứng dụng:** Bảo mật giá rẻ, mã nguồn mở hỗ trợ nghiên cứu tiếp theo.

Tài liệu tham khảo

[1] M. A. Uddin and I. H. Sarker, “An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach,” *CoRR*, vol. abs/2402.13871, 2024, doi: 10.48550/ARXIV.2402.13871.
[2] Y. Wang, W. Ma, H. Xu, Y. Liu, and P. Yin, “A Lightweight Multi-View Learning Approach for Phishing Attack Detection Using Transformer with Mixture of Experts,” *Applied Sciences*, vol. 13, no. 13, 2023, doi: 10.3390/app13137429.
[3] R. Meléndez, M. Ptaszynski, and F. Masui, “Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection,” *Electronics (Basel)*, vol. 13, no. 24, 2024, doi: 10.3390/electronics13244877.
[4] A. Al-Subaiey, M. Al-Thani, N. Abdullah Alam, K. F. Antora, A. Khandakar, and S. A. Uz Zaman, “Novel interpretable and robust web-based AI platform for phishing email detection,” *Computers and Electrical Engineering*, vol. 120, Dec. 2024, doi: 10.1016/j.compeleceng.2024.109625.
[5] N. Q. Do, A. Selamat, H. Fujita, and O. Krejcar, “An integrated model based on deep learning classifiers and pre-trained transformer for phishing URL detection,” *Future Generation Computer Systems*, vol. 161, pp. 269–285, Dec. 2024, doi: 10.1016/j.future.2024.06.031.

Tài liệu tham khảo

[6] D. M. Divakaran and A. Oest, “Phishing Detection Leveraging Machine Learning and Deep Learning: A Review,” *IEEE Secur. Priv.*, vol. 20, no. 5, pp. 86–95, 2022, doi: 10.1109/MSEC.2022.3175225.
[7] Y. Wang, W. Zhu, H. Xu, Z. Qin, K. Ren, and W. Ma, “A Large-Scale Pretrained Deep Model for Phishing URL Detection,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICASSP49357.2023.10095719.
[8] S. Asiri, Y. Xiao, and T. Li, “PhishTransformer: A Novel Approach to Detect Phishing Attacks Using URL Collection and Transformer,” *Electronics (Switzerland)*, vol. 13, no. 1, Jan. 2024, doi: 10.3390/electronics13010030.
[9] Y. Ma, G. Dobbie, and N. A. G. Arachchilage, “Combating Phishing in the Age of Fake News: A Novel Approach with Text-to-Text Transfer Transformer,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jul. 2024. doi: 10.1145/3660512.3665523.