# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://youtu.be/a0hVoal02dE
- Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/tdHuy22/CS2205.FEB2025/blob/main/Huy%20Tạ%20Duy%20-%20CS2205.FEB2025.DeCuong.FinalReport.Template.Slide.pdf

- Họ và Tên: Tạ Duy Huy
- MSSV: 240202023

- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 15
- Link Github:
  https://github.com/tdHuy22/CS2205.FEB2025.git

# ĐỀ CƯƠNG NGHIÊN CỨU

**TÊN ĐỀ TÀI (IN HOA)**

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

**TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**

DETECTING PHISHING EMAILS LEADING TO RANSOMWARE USING A LIGHTWEIGHT TRANSFORMER MODEL

**TÓM TẮT** *(Tối đa 400 từ)*

Ransomware đã trở thành mối đe dọa nghiêm trọng [1], gây thiệt hại hàng tỷ USD hàng năm thông qua các cuộc tấn công email lừa đảo tinh vi. Những tổn thất này ảnh hưởng nặng nề nhất đến người dùng cá nhân, doanh nghiệp vừa và nhỏ - những đối tượng thường thiếu khả năng đầu tư vào các giải pháp bảo mật đắt tiền. Đáp ứng thách thức này, nghiên cứu đưa ra một phương pháp kết hợp mô hình Transformer nhẹ DistillBERT [1] và Explainable-AI với LIME [1] tạo ra một giải pháp vừa minh bạch vừa có khả năng hoạt động mượt mà trên những thiết bị cấu hình khiêm tốn [2].

Dự kiến, hiệu suất của mô hình rất cao với F1-score [1] vượt ngưỡng 95%, đồng thời duy trì tỷ lệ dương tính giả ở mức thấp so với những phương pháp kinh điển như SVM [3] hay LSTM [1]. Khả năng tối ưu hóa cho phép triển khai linh hoạt trên các nền tảng từ laptop đến Raspberry Pi, đảm bảo xử lý email nhanh và mang đến giải pháp bảo mật với chi phí hợp lý. Đặc biệt, những giải thích từ LIME [1] sẽ làm nổi bật các từ khóa đáng ngờ mà dễ hiểu, hỗ trợ đắc lực cho những người dùng thiếu kiến thức chuyên môn.

Việc kết hợp Transformer nhẹ cùng XAI (Explainable-AI) [1] là ý tưởng mới trong lĩnh vực an toàn thông tin, hứa hẹn ứng dụng rộng rãi trong an ninh mạng. Hơn nữa, kết quả nghiên cứu cùng mã nguồn mở sẽ tạo tiền đề vững chắc cho các nghiên cứu tiếp theo, đặc biệt quan trọng khi mối đe dọa ransomware ngày một phức tạp và tinh vi.

**GIỚI THIỆU** *(Tối đa 1 trang A4)*

Trong bối cảnh an ninh mạng hiện tại, email lừa đảo (phishing emails) đóng vai trò là cửa ngõ chính cho ransomware - mã hóa dữ liệu và đòi tiền chuộc từ nạn nhân, trở thành mối đe doạ nghiêm trọng [1] với thiệt hại hàng tỷ USD hàng năm.Năm 2023, các cuộc tấn công lừa đảo này đã làm thiệt hai 18,7 tỉ USD dựa vào các báo cáo khác nhau [3], chủ yếu nhắm vào những mục tiêu dễ tổn thương như người dùng cá nhân và doanh nghiệp nhỏ.

Thực tế cho thấy các phương pháp phát hiện truyền thống như SVM hay Random Forest gặp hạn chế đáng kể trong việc phân tích ngữ cảnh phức tạp của email lừa đảo [3]. Mặt khác, dù các mô hình Transformer tiên tiến như BERT thể hiện tiềm năng vượt trội, chúng lại đòi hỏi tài nguyên tính toán khổng lồ [1]. Nhận thức được khoảng trống này, đề tài đề xuất một hướng tiếp cận cân bằng: sử dụng mô hình Transformer nhẹ dựa trên DistilBERT, được tối ưu hóa đặc biệt cho các thiết bị cấu hình thấp như laptop hoặc Raspberry Pi.

Nhu cầu phát hiện lừa  ở thời điểm hiện nay (2025-2026) đặc biệt ngày càng cấp thiết trong bối cảnh các cuộc tấn công ransomware gia tăng không ngừng. Các vụ việc nổi bật như WannaCry (2017) và Colonial Pipeline (2021) đã khẳng định tầm quan trọng của việc phòng chống sớm. Đồng thời, đây cũng là giai đoạn Transformer đang định hình lại xu hướng chủ đạo trong cả lĩnh vực AI và an ninh mạng [1].

Mô hình được thiết kế đặc biệt để hoạt động trên các thiết bị cấu hình thấp như laptop và Raspberry Pi, mở rộng đáng kể khả năng tiếp cận bảo vệ trước các mối đe dọa lừa đảo, tận dụng tài nguyên mã nguồn mở như Hugging Face và Google Colab.

- **Input**: Dữ liệu email từ các tập công khai (Enron Email Dataset, Kaggle phishing datasets) [4] và mẫu email ransomware từ MalwareTrafficAnalysis. Email được tiền xử lý và chuyển đổi thành đặc trưng văn bản (word embeddings) [3], sau đó xử lý mất cân bằng lớp bằng kỹ thuật oversampling/undersampling .
- **Output**: Phân loại email thành “lừa đảo” hoặc “hợp pháp” với độ chính xác cao [1], [5], kèm giải thích dự đoán bằng XAI (LIME) [1]. Mô hình nhẹ triển khai được trên thiết bị có tài nguyên hạn chế.

**MỤC TIÊU** *(Viết trong vòng 3 mục tiêu)*

**1. Phát triển mô hình phân loại email lừa đảo với độ chính xác cao**

Xây dựng một mô hình Transformer nhẹ để phân loại email thành "lừa đảo" hoặc "hợp pháp", tập trung vào các email liên quan đến ransomware, đạt hiệu suất phân loại vượt trội so với các phương pháp học máy truyền thống.

**2. Triển khai mô hình trên thiết bị cấu hình thấp**

Tối ưu hoá mô hình để triển khai trên thiết bị cấu hình thấp như laptop và Raspberry Pi thông qua kỹ thuật nén mô hình, lượng tử hoá và cắt tỉa. Mô hình cuối cùng có kích thước dưới 100MB và thời gian suy luận dưới 1 giây.

**3. Cung cấp giải thích minh bạch cho các dự đoán phân loại**

Triển khai kỹ thuật LIME [1] để cung cấp giải thích trực quan cho quyết định phân loại, làm nổi bật các từ khóa và đặc trưng quan trọng ảnh hưởng đến dự đoán. Hệ thống giải thích được thiết kế để người dùng không chuyên dễ hiểu, với mục tiêu đạt điểm đánh giá 4/5 về độ dễ hiểu.

## NỘI DUNG VÀ PHƯƠNG PHÁP

**Thu thập và tiền xử lý dữ liệu:**

- Dữ liệu được thu thập từ ba nguồn chính: Enron Email Dataset (500.000 email hợp pháp) [4], Phishing Email Dataset từ Kaggle (10.000 email lừa đảo) [4], và 1.000 email ransomware từ MalwareTrafficAnalysis liên quan đến các chiến dịch Ryuk và WannaCry.
- Tiền xử lý bao gồm loại bỏ email trùng lặp [1], xóa HTML và tệp đính kèm bằng BeautifulSoup, lọc ký tự không liên quan (emoji, mã đặc biệt) [4]. Dữ liệu sau xử lý được mã hóa bằng BERT tokenizer (Hugging Face), giới hạn tối đa 512 token để tạo input embeddings tương thích với DistilBERT [3].
- Do mất cân bằng lớp (ransomware và phishing chiếm tỷ lệ nhỏ) [1], [5], kỹ thuật *class weighting* được áp dụng trong hàm mất mát để tăng trọng số cho lớp thiểu số, giúp mô hình học hiệu quả hơn mà không làm thay đổi phân phối dữ liệu.

**Huấn luyện và đánh giá mô hình:**

- Mô hình được huấn luyện dựa trên DistilBERT – phiên bản rút gọn của BERT với khoảng 66 triệu tham số – nhờ khả năng cân bằng hiệu suất và tốc độ [1], [3]. Quá trình fine-tune được thực hiện trên Google Colab với GPU Tesla T4. Tham số huấn luyện gồm: batch size = 16, learning rate = 2e-5, tối đa 5 epoch, sử dụng Cross-Entropy loss và AdamW optimizer [1], [3], [5].
- Tập dữ liệu được chia theo tỷ lệ 80:10:10 cho huấn luyện, xác thực, và kiểm tra. Đánh giá tập trung vào F1-score (mục tiêu > 95%), cùng với accuracy và false positive rate (FPR < 1%) để đảm bảo tính chính xác và an toàn trong phân loại email.

**Tối ưu hoá và triển khai:**

- Để tối ưu hóa DistilBERT [1] cho các thiết bị cấu hình thấp, ba kỹ thuật chính đã được áp dụng. Quantization (float32 → int8) bằng Hugging Face Optimum, giúp nén kích thước mô hình xuống dưới 100MB và RAM yêu cầu dưới 8GB. Pruning [2] loại bỏ 20% các trọng số ít quan trọng, nhằm giảm thời gian suy luận (inference time) mỗi email xuống dưới 1 giây. Sau đó, mô hình được chuyển đổi sang định dạng ONNX thông qua ONNX Runtime để tối ưu hóa tốc độ xử lý trên cả CPU và kiến trúc ARM.
- Mô hình được thử nghiệm trên laptop cấu hình thấp (Windows/Linux, Intel/AMD, RAM 8GB) và Raspberry Pi 4 (RAM 4GB). Việc triển khai thông qua ứng dụng web Flask cho phép người dùng nhập nội dung email và nhận kết quả phân loại thời gian thực. Quá trình inference sử dụng ONNX Runtime kết hợp BERT tokenizer, đồng bộ với pipeline huấn luyện.

**Tích hợp khả năng giải thích XAI (Explainable-AI):**

- Kỹ thuật tích hợp: LIME [1] được áp dụng để xác định các từ hoặc cụm từ ảnh hưởng mạnh đến quyết định phân loại (ví dụ: *"urgent"*, *"payment"*). Trọng số đóng góp của từng từ được tính toán dựa trên thay đổi xác suất dự đoán khi từ đó bị loại bỏ hoặc thêm vào. Kết quả được hiển thị trực quan trên giao diện Flask

dưới dạng danh sách từ khóa kèm trọng số (ví dụ: *urgent: 0.75*), hỗ trợ người dùng hiểu lý do email bị gắn nhãn.

- Đánh giá tính minh bạch: Mô hình được thử nghiệm trên 100 email từ tập kiểm tra, ghi nhận các từ khóa LIME xác định là quan trọng. Một khảo sát người dùng (10 – 15 người dùng không chuyên) được tiến hành để đánh giá mức độ dễ hiểu của giải thích, sử dụng thang điểm Likert từ 1 đến 5. Mục tiêu là đạt điểm trung bình trên 4, cho thấy giải thích từ mô hình là rõ ràng và có thể tiếp cận với người dùng phổ thông.

## KẾT QUẢ MONG ĐỢI

Đề tài hướng đến giải pháp phát hiện email lừa đảo dẫn đến ransomware với các tiêu chí: hiệu quả, nhẹ, và minh bạch. Mô hình DistilBERT [1] được tinh chỉnh nhằm đạt F1-score trên 95%, vượt trội so với các mô hình truyền thống như SVM và LSTM [3], qua đó nâng cao khả năng phát hiện sớm các mối đe dọa qua email.

Giải pháp được tối ưu để triển khai trên thiết bị có cấu hình hạn chế (RAM 8GB hoặc thấp hơn), bao gồm cả Raspberry Pi 4, với mô hình có kích thước dưới 100MB và thời gian xử lý mỗi email dưới 1 giây, phù hợp cho các hệ thống bảo mật giá rẻ.

Việc tích hợp LIME [1] cho phép giải thích các dự đoán một cách trực quan, giúp người dùng hiểu rõ lý do email bị phân loại là nguy hiểm. Đánh giá qua khảo sát người dùng dự kiến mức độ dễ hiểu đạt trên 4/5, góp phần tăng cường tính minh bạch và độ tin cậy của hệ thống.

## TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*


[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.

[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.