THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/UaBSMEckOgs
  (ví dụ: https://www.youtube.com/watch?v=AWq7uw-36Ng)
- Link slides (dạng .pdf đặt trên Github):
  https://github.com/hongphuoc0209/CS2205.MAR2024
  (ví dụ: https://github.com/mynameuit/CS2205.APR2023/TenDeTai.pdf)
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Nguyễn Cẩm Hồng Phước • MSSV: 230201024 	• Lớp: CS2205.MAR2024 • Tự đánh giá (điểm tổng kết môn): 9/10 • Số buổi vắng: 1 • Số câu hỏi QT cá nhân: 3 • Link Github:

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN TẬP TIN PDF ĐỘC HẠI SỬ DỤNG MÔ HÌNH TRANSFORMER VÀ TRÍ TUỆ NHÂN TẠO KHẢ DIỄN GIẢI

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

MALICIOUS PDF FILES DETECTION USING TRANSFORMER AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

TÓM TẮT *(Tối đa 400 từ)* ***[Số từ hiện tại: 350 từ]***

Tập tin PDF là loại tập tin được sử dụng phổ biến trong các loại tập tin được đính kèm qua thư điện tử. Việc đính kèm mã độc vào tập tin PDF ngày càng phổ biến. Có nhiều nghiên cứu trong việc phát hiện tập tin PDF. Tuy nhiên, cần thiết phải có các nghiên cứu sử dụng các hướng tiếp cận mới và hiện đại hơn cũng như bổ sung các bộ dataset để tăng tính hiệu quả trong việc huấn luyện các mô hình. Luận văn này tập trung vào việc đề xuất hệ thống phát hiện tập tin PDF độc hại sử dụng mô hình Transformer và trí tuệ nhân tạo khả diễn giải (XAI). Nghiên cứu này sử dụng các mô hình xử lý ngôn ngữ tự nhiên để vector hóa các đặc trưng được rút trích từ các tập tin PDF. Các biến thể của mô hình Transformer được sử dụng để xây dựng mô hình phát hiện tập tin PDF độc hại. Bên cạnh đó, luận văn áp dụng các nền tảng trí tuệ nhân tạo khả diễn giải như SHAP và LIME để giải thích sự đóng góp của các đặc trưng đối với mô hình phát hiện tập tin PDF độc hại. Ngoài ra, để tăng cường số lượng mẫu thử cho tập dữ liệu thử nghiệm hiện tại, luận văn áp dụng mạng sinh đối kháng (GAN - Generative Adversarial Network) để phát sinh các mẫu thử là tập tin PDF. Bộ dữ liệu thử nghiệm mở rộng này được sử dụng để huấn luyện lại các mô hình nhằm giúp tăng độ chính xác cho các mô hình khi đánh giá chúng với nhiều mẫu thử hơn ngoài thực tế. Các đóng góp chính của luận văn bao gồm việc vector hóa các đặc trưng bằng mô hình xử lý ngôn ngữ tự nhiên, bổ sung mẫu thử cho bộ dữ liệu thử nghiệm bằng GAN, huấn luyện và tái huấn luyện mô hình để tạo mô hình chính xác hơn, giải thích sự

đóng góp của các đặc trưng đối với mô hình bằng XAI.

GIỚI THIỆU *(Tối đa 1 trang A4)* ***[Độ dài hiện tại: 1 trang A4]***

Tệp PDF được phát triển bởi Adobe Systems để hiển thị tài liệu điện tử, độc lập với phần mềm, phần cứng hoặc hệ điều hành. Tệp PDF có bốn thành phần chính. Header, Body, Cross-Reference Table, Trailer. Việc nắm rõ cấu tạo của tập tin PDF có thể giúp ích trong việc xây dựng các công cụ rút trích đặc trưng của loại tập tin này phục vụ việc huấn luyện mô hình phát hiện tập tin PDF độc hại.

Có nhiều nghiên cứu liên quan đến việc xây dựng mô hình phát hiện tập tin PDF độc hại. Các mô hình thường được sử dụng như CNN [1], Random Forest [2], SVM [3], XGBoost [4], AdaBoost [4],… Tuy nhiên, đa phần các nghiên cứu này sử dụng cách rút trích đặc trưng bằng cách điểm danh sự xuất hiện của các loại đặc trưng trong tập tin PDF. Trong khi đó, tập tin PDF có thể chứa các loại thông tin dạng chuỗi trong các script đính kèm. Do đó, việc áp dụng các mô hình xử lý ngôn ngữ tự nhiên để xử lý các loại đặc trưng dạng chuỗi này dự kiến mang lại hiệu quả trong việc vector hóa các đặc trưng của tập tin PDF. Đây là một trong những mục tiêu của luận văn.

Các nghiên cứu hiện tại chủ yếu sử dụng dataset CIC-Evasive-PDFMal2022 dataset [5]. Dataset này chứa 10.025 mẫu thử. Tuy nhiên, để xây dựng mô hình phát hiện tập tin PDF hiệu quả hơn khi đánh giá các tập tin ngoài thực tế, dataset này cần được mở rộng cả về số lượng mẫu và tính đa dạng khi phân bố các loại đặc trưng. Do đó, việc ứng dụng mạng sinh đối kháng (GAN) [6] để bổ sung các các mẫu thử cho bộ dataset hiện tại là cần thiết. Đây là một trong những mục tiêu của luận văn.

Các nghiên cứu hiện tại chủ yếu đề xuất mô hình phát hiện tập tin độc hại bằng cách sử dụng các mô hình học máy/học sâu phổ biến như, CNN, Random Forest, SVM,… Tuy nhiên, họ chưa chú tâm nhiều trong việc giải thích các mô hình. Cụ thể là giải thích sự đóng góp của các đặc trưng đối với mô hình. Việc hiểu rõ sự đóng góp này giúp ích nhiều trong việc lựa chọn đặc trưng tối ưu. Đặc biệt trong ngữ cảnh triển

khai các mô hình trên điện thoại di động, nơi có tài nguyên tính toán thấp. Luận văn này sử dụng các nền tảng trí tuệ nhân tạo khả diễn giải (XAI) để giải thích các mô hình.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Xây dựng được mô-đun rút trích đặc trưng bằng mô hình xử lý ngôn ngữ tự nhiên để rút trích đặc trưng cho việc xây dựng mô hình phát hiện tập tin PDF độc hại sử dụng các biến thể của Transformer.
- Xây dựng được mô-đun tạo mẫu thử bổ sung cho dataset hiện tại bằng cách sử dụng mạng sinh đối kháng như DCGAN.
- Đánh giá độ chính xác của mô hình bằng bộ dữ liệu thử nghiệm CIC-Evasive-PDFMal2022 và bộ dữ liệu được tạo từ GAN trong luận văn. Giải thích đóng góp của các đặc trưng với mô hình bằng một trong những nền tảng XAI như SHAP, LIME.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Nội dung nghiên cứu 1: Xây dựng mô-đun rút trích đặc trưng bằng mô hình xử lý ngôn ngữ tự nhiên để rút trích đặc trưng cho việc xây dựng mô hình phát hiện tập tin PDF độc hại sử dụng các biến thể của Transformer.

Phương pháp nghiên cứu:

- Nghiên cứu cấu trúc tập tin PDF và cách rút trích các thông tin từ tập tin PDF qua thư viện PyMuPDF[7].
- Khảo sát và tổng hợp các công trình nghiên cứu liên quan (literature review) về các phương pháp phát hiện tập tin PDF độc hại để xác định các khoảng trống nghiên cứu (research gaps).
- Nghiên cứu cách thức vector hóa các đặc trưng bằng các mô hình xử lý ngôn

ngữ tự nhiên như word2vec, BERT, DistilBERT, RoBERT, XLNet[8].

- Thử nghiệm các cách thức vector hóa khác nhau, đánh giá về thời gian, bộ nhớ cần thiết cho từng mô hình.

Nội dung nghiên cứu 2: Xây dựng mô-đun tạo mẫu thử bổ sung cho dataset hiện tại bằng cách sử dụng mạng sinh đối kháng như DCGAN**.**

Phương pháp nghiên cứu:

- Nghiên cứu và triển khai mô hình mạng sinh đối kháng phổ biến như DCGAN.

- Từ các vector được rút trích từ nội dung 1, thực hiện tạo vector tùy chỉnh bằng GAN.

- Từ vector tùy chỉnh này thực hiện tạo tập tin PDF tùy chỉnh từ GAN.

- Kiểm thử chất lượng mẫu thử được tạo từ GAN.

Nội dung nghiên cứu 3: Đánh giá độ chính xác của mô hình bằng bộ dữ liệu thử nghiệm CIC-Evasive-PDFMal2022 và bộ dữ liệu được tạo từ GAN trong luận văn. Giải thích đóng góp của các đặc trưng với mô hình bằng một trong những nền tảng XAI như SHAP, LIME**.**

Phương pháp nghiên cứu:

- Sử dụng các độ đo như Accuracy, F1-Score, Recall, Precision để đánh giá mô hình được huấn luyện từ dataset gốc và dataset được tạo từ GAN.

- Cài đặt các nền tảng trí tuệ nhân tạo khả diễn giải như SHAP, LIME [9] để giải thích các mô hình được huấn luyện.

- Biện luận các kết quả thử nghiệm.

- Viết báo cáo khoa học, tham gia hội nghị, viết luận văn, báo cáo thử.

Kế hoạch thực hiện:

Vì luận văn theo hình thức hướng nghiên cứu (15 tín chỉ), nên thời gian thực hiện theo quy định là 12 tháng. Kế hoạch thực hiện các nội dung chính như sơ đồ Gantt sau. Tuy nhiên, học viên và giảng viên hướng dẫn đặt mục tiêu hoàn thành luận văn sớm hơn kế hoạch 12 tháng.

Công việc	Tháng 1	Tháng 2	Tháng 3	Tháng 4	Tháng 5	Tháng 6	Tháng 7	Tháng 8	Tháng 9	Tháng 10	Tháng 11	Tháng 12
ND 1												
ND 2												
ND 3												

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

Các kết quả mong đợi của luận văn bao gồm:

- Xây dựng được hệ thống phát hiện tập tin PDF độc hại bằng cách sử dụng các biến thể mô hình Transformer và có khả năng giải thích được sự đóng góp của các đặc trưng đến mô hình.
- Xây dựng được mô-đun phát sinh mẫu thử bằng cách dùng mạng sinh đối kháng.
- Tích hợp nền tảng trí tuệ nhân tạo khả diễn giải trong bài toán phát hiện tập tin PDF độc hại bằng học sâu.

Công bố 01 báo cáo khoa học trong hội nghị quốc tế có kỷ yếu được chỉ mục trong Scopus.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges,

applications, future directions," *Journal of big Data,* vol. 8, pp. 1-74, 2021.

[2] J. Hatwell, M. M. Gaber, and R. M. A. Azad, "CHIRPS: Explaining random forest classification," *Artificial Intelligence Review,* vol. 53, pp. 5747-5788, 2020.

[3] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing,* vol. 408, pp. 189-215, 2020.

[4] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access,* vol. 10, pp. 99129-99149, 2022.

[5] M. Issakhani, P. Victor, A. Tekeoglu, and A. H. Lashkari, "PDF Malware Detection based on Stacking Learning," in *ICISSP*, 2022, pp. 562-570.

[6] F. Zhong, X. Cheng, D. Yu, B. Gong, S. Song, and J. Yu, "MalFox: Camouflaged adversarial malware example generation based on conv-GANs against black-box detectors," *IEEE Transactions on Computers,* 2023.

[7] Artifex. (2024, May 10). *PyMuPDF 1.24.3 documentation*. Available: https://pymupdf.readthedocs.io/en/latest/

[8] K. Giapantzis and S. T. Halkidis, "XLCNN: A Transformer Model for Malware Detection," *International Journal of Computer Science and Information Security* vol. 21, no. 7, 2023.

[9] S. Lundberg. (2023, October, 10). *SHapley Additive exPlanations*. Available: https://shap.readthedocs.io/en/latest/