ĐỀ CƯƠNG ĐỀ TÀI ĐỒ ÁN CHUYÊN NGÀNH

I. THÔNG TIN CHUNG.

1. Sinh viên/Nhóm thực hiện:

Tên : Lê Quang Thịnh	MSSV:2280603087	Lớp:22DTHB3
Tên : Nguyễn Nhật Hào	Mssv:2280618857	Lớp:22DTHB1

2. **GVHD:** GV . Đinh Phương Nam

3. Tên đề tài: Ứng dụng Machine Learning trong việc phòng chống phát hiện mã độc ẩn trong file PDF / Office

II. NỘI DUNG ĐỀ CƯƠNG.

1. Giới thiệu và cơ sở lý thuyết.

Trong thời đại công nghệ số, các tệp tin văn phòng (Office: Word, Excel, PowerPoint) và PDF được sử dụng phổ biến trong môi trường học tập, doanh nghiệp và chính phủ. Tuy nhiên, chính sự phổ biến này đã khiến chúng trở thành mục tiêu hấp dẫn cho tin tặc khai thác. Các tệp PDF/Office thường được sử dụng làm công cụ phát tán mã độc thông qua email lừa đảo (phishing), tấn công có chủ đích (APT) hoặc các chiến dịch spam.

Với sự phổ biến của tệp PDF trong việc chia sẻ thông tin và tài liệu, các tin tặc đã tận dụng điều này để nhắm mục tiêu tấn công người dùng. Những tệp PDF có thể được chèn mã độc ngầm, hoặc các liên kết và điều hướng người dùng đến các trang web nguy hiểm [1].

Chính vì vậy, việc nghiên cứu và xây dựng hệ thống phát hiện mã độc ẩn trong file PDF/Office là cần thiết, góp phần nâng cao khả năng phòng thủ trước các mối đe dọa an ninh mạng.

1.1. Bối cảnh nghiên cứu.

PDF, loại tệp tin được sử dụng phổ biến trong môi trường kinh doanh, đang trở thành một mục tiêu hấp dẫn cho các cuộc tấn công mạng. Điều đáng chú ý là nạn nhân thường thiếu khả năng phòng ngừa với loại tệp tin thông dụng này, so với các tệp tin như EXE mà họ có thể cảnh giác hơn. Đáng lo ngại hơn, nhiều người thậm chí không nhận ra rằng PDF cũng có thể được sử dụng để thực hiện các hành vi xấu.

Virus, Trojan, mã độc có nhiều cách ẩn bên trong tệp PDF và thường được phát tán thông qua các tệp được tải xuống hoặc tệp đính kèm trong email như ebook hoặc các tài liệu khác. Chúng thường được gửi từ các nguồn không xác định hoặc không quen thuộc.

Theo Palo Alto Networks, *"Phần lớn mọi người đều biết rằng không nên nhấp vào các liên kết lạ trong email, do đó tin tặc đã chuyển sang các kế hoạch lừa đảo thông qua sử dụng định dạng PDF, vì nó có thể hiệu quả hơn so với email thông thường chỉ chứa một liên kết đơn giản dễ nhận biết"*.
Đặc biệt, tin tặc thường có kỹ năng trong các chiêu trò kỹ thuật xã hội, do đó các email lừa đảo thường được thiết kế để thao túng cảm xúc của người đọc hoặc đánh vào những thành kiến vô thức của họ.
Theo đó, một số kế hoạch lừa đảo thông qua tệp PDF đã được tin tặc sử dụng phổ biến. Đặc điểm chung của những kế hoạch này là sử dụng các quy ước đặt tên mà người dùng thường gặp trong môi trường kinh doanh [2].

1.2. Vấn đề file PDF/Ofice khi chứa mã độc.

Tệp tin PDF và Office vốn được xem là định dạng tài liệu phổ biến, đáng tin cậy, được sử dụng rộng rãi trong các hoạt động văn phòng, hành chính, học tập và giao dịch trực tuyến. Tuy nhiên, chính vì mức độ phổ biến và sự tin tưởng này mà chúng trở thành mục tiêu khai thác lý tưởng của tin tặc. Việc chèn mã độc vào các tệp tài liệu không chỉ gây khó khăn cho người dùng trong việc nhận diện, mà còn thách thức các giải pháp bảo mật truyền thống [3].

1.3. Tính cấp thiết của đề tài.

Theo báo cáo của Kaspersky (2023), có đến hơn 40% mã độc được phát tán qua email đính kèm file Office hoặc PDF. Các biến thể mới liên tục xuất hiện, áp dụng kỹ thuật mã hóa, làm rối, hoặc chống phân tích khiến giải pháp truyền thống gần như bất lực.

Báo cáo Nghiên cứu Xu hướng mối đe dọa mạng của Unit 42 (nhóm nghiên cứu về thông tin mối đe dọa toàn cầu) của Palo Alto Networks cũng cho biết số vụ tấn công bằng phần mềm độc hại trung bình mà mỗi tổ chức trong ngành sản xuất, tiện ích và năng lượng gặp phải đã tăng 238% (từ năm 2021 đến năm 2022). Điều này cho thấy sự gia tăng đáng kể của mối đe dọa này và sự cần thiết của việc thực hiện các biện pháp bảo mật hiệu quả để đối phó với chúng [4].

1.4. Mục tiêu nghiên cứu.

Đề tài "Phát hiện mã độc ẩn trong file PDF/Office" hướng tới việc xây dựng cơ sở lý thuyết và mô hình ứng dụng thực tiễn nhằm phát hiện và ngăn chặn các mối đe dọa an ninh mạng xuất phát từ tài liệu văn phòng. Cụ thể, các mục tiêu chính của nghiên cứu bao gồm:

1.4.1. Mục tiêu tổng quát

Nghiên cứu, phân tích và đề xuất một phương pháp phát hiện mã độc ẩn trong file PDF/Office dựa trên sự kết hợp giữa phân tích đặc trưng và học máy (Machine Learning).

Xây dựng hệ thống thử nghiệm có khả năng nhận diện tài liệu độc hại, từ đó góp phần nâng cao an toàn thông tin cho người dùng cá nhân và tổ chức.

1.4.2. Mục tiêu cụ thể

1. Khảo sát tình hình và kỹ thuật tấn công

Tìm hiểu cách thức tin tặc chèn mã độc vào file PDF/Office thông qua macro, JavaScript, OLE Object, hoặc khai thác lỗ hổng zero-day.

Phân tích các trường hợp điển hình để rút ra đặc điểm nhận dạng.

2. Xác định và trích xuất đặc trưng (feature extraction)

Đề xuất các đặc trưng tĩnh (cấu trúc file, macro, metadata, script nhúng).

Khai thác thêm các đặc trưng động (hành vi thực thi khi mở file).

3. Ứng dụng học máy vào phát hiện mã độc

Thử nghiệm các thuật toán Machine Learning (Decision Tree, Random Forest, SVM, Logistic Regression).

Đánh giá hiệu quả của các mô hình dựa trên độ chính xác, độ nhạy (recall), độ đặc hiệu (precision) và F1-score.

4. Xây dựng và kiểm thử hệ thống mẫu (prototype)

Xây dựng công cụ thử nghiệm có khả năng tự động phân tích file PDF/Office.

Kiểm thử với bộ dữ liệu thực tế để đánh giá hiệu quả.

5. Đề xuất hướng phát triển trong tương lai

Nâng cấp mô hình bằng cách áp dụng Deep Learning.

Xem xét triển khai thực tế trong môi trường doanh nghiệp hoặc dịch vụ đám mây.

2. Thiết kế và triển khai hệ thống.

Đề tài "Phát hiện mã độc ẩn trong file PDF/Office" được triển khai theo các nội dung nghiên cứu chính sau:

2.1. Khảo sát và phân tích lý thuyết

Tìm hiểu về các loại mã độc thường ẩn trong tài liệu PDF/Office.

Phân tích kỹ thuật chèn mã độc:

Office: Macro VBA, OLE Object, Dynamic Data Exchange (DDE).

PDF: JavaScript nhúng, Embedded Object, khai thác lỗ hổng zero-day.

Đánh giá ưu, nhược điểm của các phương pháp phát hiện truyền thống: quét chữ ký, sandbox phân tích động, heuristic detection.

2.2. Thu thập và xây dựng dữ liệu nghiên cứu

Thu thập file PDF/Office từ nhiều nguồn:

Benign files: tài liệu văn phòng sạch từ người dùng thực tế, tài liệu học tập, báo cáo chính thống.

Malicious files: từ các kho dữ liệu như VirusShare, MalwareBazaar, Contagio.

Tiến hành gán nhãn dữ liệu (Benign/Malicious).

Chuẩn hóa định dạng dữ liệu để dễ dàng xử lý.

2.3. Trích xuất đặc trưng (Feature Extraction)

Đặc trưng tĩnh (Static Features):

Thông tin metadata: tên tác giả, ngày tạo, kích thước file.

Số lượng macro/JavaScript nhúng.

Sự tồn tại của OLE Object, external link.

Cấu trúc file: header, objects, stream.

Đặc trưng động (Dynamic Features):

Tiến trình được tạo ra khi mở file.

Hành vi kết nối mạng (DNS, HTTP, IP).

Thay đổi hệ thống (registry, file hệ thống).

Biểu diễn dữ liệu:

Lưu trữ đặc trưng ở dạng vector số hóa (CSV/JSON).

Tiến hành chuẩn hóa (normalization) và loại bỏ nhiễu.

2.4. Lựa chọn và xử lý đặc trưng

Ứng dụng kỹ thuật thống kê (Chi-square, Information Gain) để chọn đặc trưng quan trọng.

Giảm chiều dữ liệu để tối ưu tốc độ huấn luyện.

Xây dựng tập dữ liệu huấn luyện và kiểm thử cân bằng.

2.5. Ứng dụng Machine Learning trong phát hiện

Lựa chọn các thuật toán:

Decision Tree.

Random Forest.

Support Vector Machine (SVM).

Logistic Regression.

Huấn luyện mô hình bằng tập dữ liệu đã chuẩn bị.

So sánh hiệu quả các mô hình dựa trên các chỉ số: Accuracy, Precision, Recall, F1-score, ROC-AUC.

2.6. Xây dựng hệ thống thử nghiệm (Prototype)

Thiết kế pipeline phát hiện mã độc:

1. Input: file PDF/Office.
2. Trích xuất đặc trưng.
3. Phân loại bằng mô hình Machine Learning.
4. Xuất kết quả: Benign hoặc Malicious.

Giao diện đơn giản cho phép người dùng chọn file để kiểm tra.

2.7. Đánh giá kết quả và phân tích

Kiểm thử hệ thống với tập dữ liệu thực tế chưa được dùng trong huấn luyện.

So sánh kết quả giữa mô hình nghiên cứu và giải pháp truyền thống.

Đánh giá ưu điểm, hạn chế và độ tin cậy.

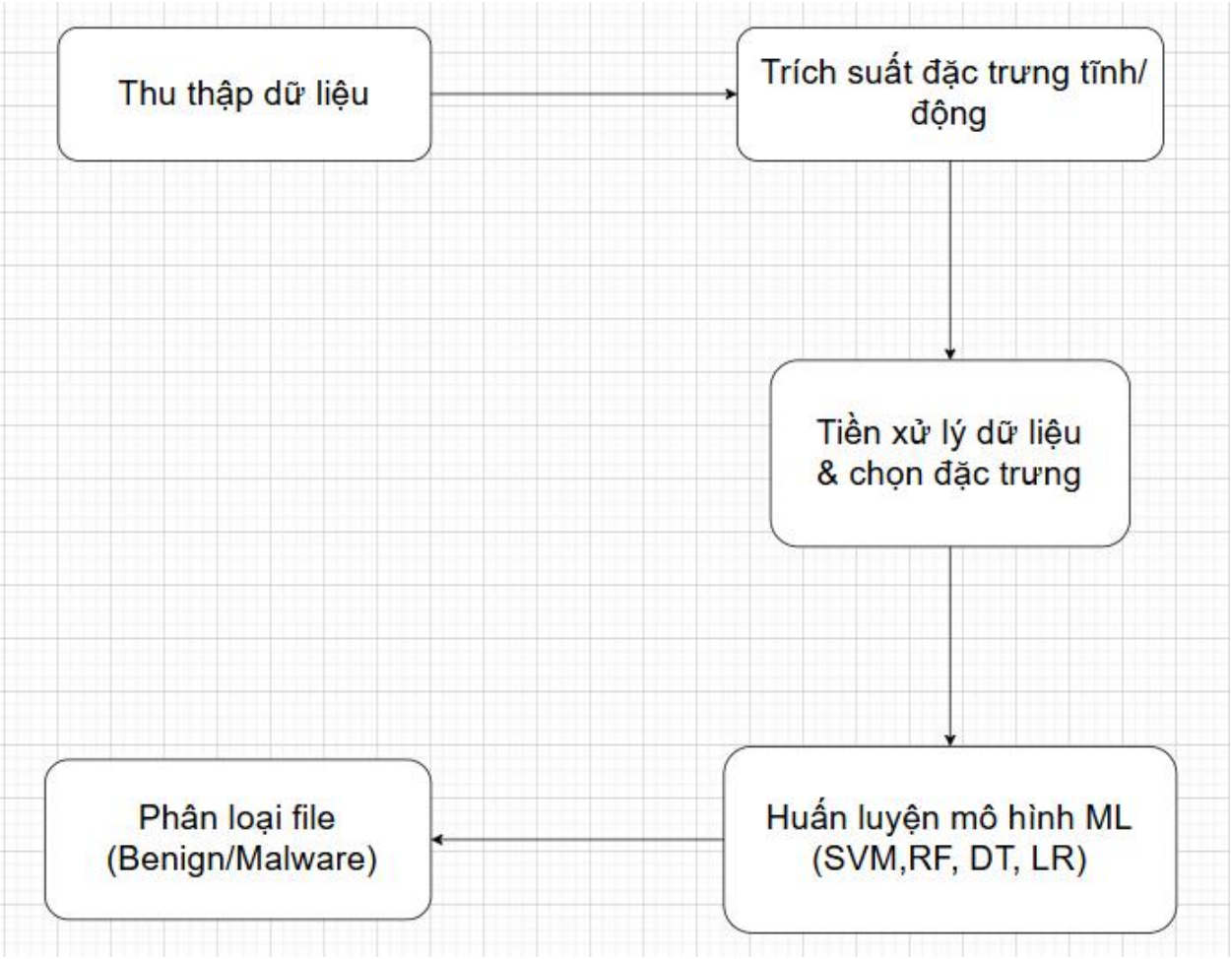
2.8. Đề xuất và hướng phát triển

Đề xuất áp dụng Deep Learning (CNN, LSTM) để cải thiện độ chính xác.

Mở rộng hệ thống để phát hiện nhiều loại tài liệu khác (PowerPoint, RTF).

Nghiên cứu triển khai trên môi trường doanh nghiệp hoặc tích hợp với email server.

2.9. Mô hình hoạt động của đề tài nghiên cứu.



3. Đánh giá và kết luận.

3.1. Tiêu chí đánh giá.

Để đánh giá hiệu quả của hệ thống phát hiện mã độc ẩn trong file PDF/Office, đề tài sử dụng các tiêu chí sau:

- Độ chính xác (Accuracy): tỷ lệ mẫu được phân loại đúng trên tổng số mẫu.
- Độ nhạy (Recall): khả năng phát hiện đúng các file chứa mã độc.
- Độ đặc hiệu/Độ chính xác (Precision): tỷ lệ mẫu được gán nhãn mã độc thực sự là mã độc.
- F1-score: chỉ số cân bằng giữa Precision và Recall, đặc biệt quan trọng trong bối cảnh dữ liệu mất cân bằng.
- ROC-AUC: diện tích dưới đường cong ROC, phản ánh khả năng phân biệt giữa file benign và malicious.
- Hiệu năng xử lý: thời gian trung bình để phân tích một file.
- Tính ổn định: khả năng xử lý file với định dạng hoặc biến thể khác nhau mà không gây lỗi.

3.2. Kết quả dự kiến.

Hệ thống đạt độ chính xác trên 90%, với Recall > 85% nhằm giảm thiểu tình trạng bỏ sót file độc hại.

Các thuật toán Random Forest và SVM được kỳ vọng cho hiệu quả tốt nhất trong việc phân loại file.

Thời gian phân tích file bằng phương pháp trích xuất đặc trưng tĩnh duy trì dưới 10 giây/file, đảm bảo tính khả dụng thực tế.

Xây dựng được prototype hệ thống có thể kiểm tra file PDF/Office và trả kết quả phân loại (Benign/Malicious) một cách trực quan.

3.3. Phân tích ưu – nhược điểm.

Ưu điểm:

Sử dụng Machine Learning giúp phát hiện được cả các mẫu mã độc chưa có chữ ký nhận diện (zero-day).

Kết hợp đặc trưng tĩnh và động, cho phép hệ thống có góc nhìn toàn diện hơn.

Hệ thống có khả năng mở rộng, có thể bổ sung thêm dữ liệu hoặc huấn luyện lại mô hình để cải thiện độ chính xác.

Nhược điểm:

Phân tích động tốn nhiều thời gian và tài nguyên, khó áp dụng cho số lượng lớn file trong thời gian ngắn.

Hệ thống phụ thuộc vào chất lượng và độ đa dạng của tập dữ liệu; nếu dữ liệu hạn chế, mô hình dễ bị overfitting.

Các kỹ thuật làm rối (obfuscation) hoặc mã độc tinh vi có thể che giấu hành vi, làm giảm hiệu quả phát hiện.

3.4. Khả năng ứng dụng thực tế.

Trong doanh nghiệp: Tích hợp vào hệ thống email gateway để tự động kiểm tra file đính kèm trước khi đến người dùng.

Trong tổ chức giáo dục/nghiên cứu: Hỗ trợ phân tích, nghiên cứu mã độc, phục vụ công tác đào tạo an ninh mạng.

Trong cá nhân/người dùng cuối: Phát triển thành công cụ kiểm tra tài liệu trước khi mở, tăng cường bảo mật khi nhận file từ nguồn không tin cậy.

Trong môi trường đám mây: Tích hợp với hệ thống lưu trữ (Google Drive, OneDrive) để phát hiện sớm mã độc trước khi lây lan.

3.5. Kết luận và hướng phát triển.

Kết luận:

Đề tài đã nghiên cứu và xây dựng mô hình phát hiện mã độc trong file PDF/Office thông qua việc trích xuất đặc trưng và ứng dụng Machine Learning. Kết quả thử nghiệm cho thấy tính khả thi và hiệu quả cao trong việc phát hiện các mẫu độc hại.

Hướng phát triển:

Áp dụng Deep Learning (CNN, RNN, Transformer) để nâng cao độ chính xác và khả năng nhận diện mã độc phức tạp.

Mở rộng tập dữ liệu đa dạng và lớn hơn, bao gồm nhiều loại file văn phòng khác (PowerPoint, RTF, ODT).

Nghiên cứu kỹ thuật phát hiện mã độc được làm rối hoặc mã hóa để tăng độ bền vững của hệ thống.

Tích hợp hệ thống vào sản phẩm thực tế (plugin email, cổng bảo mật doanh nghiệp, dịch vụ API đám mây).

4. **Kế hoạch thực hiện**

Nội dung/ Tuần	1	2	3	4	5	6	7	8	9	10
Thu thập dữ liệu PDF/Office chứa mã độc		X								
Trích xuất đặc trưng của từng dữ liệu				X						
Tiền xử lý dữ liệu và chọn đặc trưng						X				
Huấn luyện mô hình							X			
Chạy thử nghiệm và áp dụng thực tế								X		
Chỉnh sửa những lỗi cần khắc phục									X	
Ghi nhận kết quả và báo cáo										X

TÀI LIỆU THAM KHẢO:

[1]. Adobe. (n.d.). *Can PDFs contain viruses?* Adobe Acrobat. Truy cập ngày 28 tháng 9, 2025, từ https://www.adobe.com/acrobat/resources/can-pdfs-contain-viruses.html

[2]. NAIT. (2023). *Phần mềm độc hại trong tệp PDF – rủi ro lớn tiềm ẩn của doanh nghiệp*. Trung tâm An toàn thông tin NAIT. Truy cập ngày 28 tháng 9, 2025, từ https://www.nait.vn/attt/phan-mem-doc-hai-trong-tep-pdf-rui-ro-lon-tiep-cua-doanh-nghiep-375.html

[3]. Thế Giới Di Động. (2019). *Cảnh báo mã độc giả mạo file PDF và hướng dẫn phòng tránh*. Truy cập ngày 28 tháng 9, 2025, từ https://www.thegioididong.com/tin-tuc/canh-bao-ma-doc-gia-mao-file-pdf-va-huong-dan-phong-tranh-1105358

[4]. NAIT. (2023). *Phần mềm độc hại trong tệp PDF – rủi ro lớn tiềm ẩn của doanh nghiệp* (lặp lại). Trung tâm An toàn thông tin NAIT. Truy cập ngày 28 tháng 9, 2025, từ https://www.nait.vn/attt/phan-mem-doc-hai-trong-tep-pdf-rui-ro-lon-tiep-cua-doanh-nghiep-375.html

---HẾT---