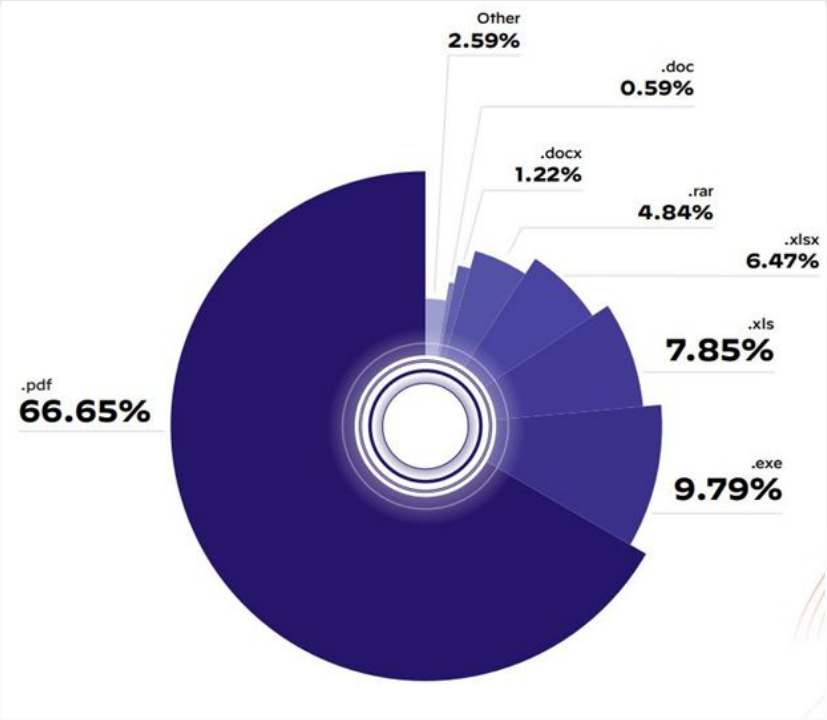
PHÁT HIỆN TẬP TIN PDF ĐỘC HẠI SỬ DỤNG MÔ HÌNH TRANSFORMER VÀ TRÍ TUỆ NHÂN TẠO KHẢ DIỄN GIẢI

Nguyễn Cẩm Hồng Phước - 230201024

Tóm tắt

- Lớp: CS2205.MAR2024
- Link Github: https://github.com/hongphuoc0209/CS2205.MAR2024
- Link YouTube video: https://youtu.be/UaBSMEckOgs
- Ảnh + Họ và Tên: Nguyễn Cẩm Hồng Phước
- Tổng số slides không vượt quá 10

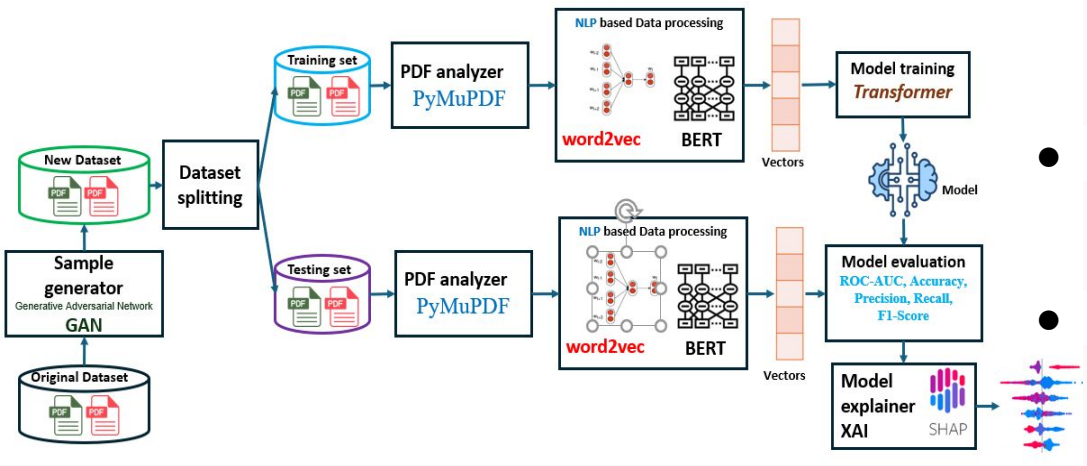
Giới thiệu



Phân bố theo định dạng tệp độc hại đính kèm qua email
(Palo Alto Network, 2023)

- Các cuộc tấn công thông qua PDF ngày càng phổ biến.
- Các nghiên cứu hiện nay sử dụng các mô hình học máy dựa trên các đặc trưng.
- Các biến thể độc hại ngày càng tinh vi, xuất hiện mới hằng ngày, dẫn đến nhu cầu về nguồn dữ liệu huấn luyện càng cao.
- Nhu cầu về tính minh bạch trong cách thức ra quyết định, phân tích các đặc trưng để cải thiện hiệu quả và hiệu suất mô hình ngày càng cao.

Mục tiêu



- Xây dựng mô hình phát hiện tập tin PDF độc hại sử dụng các biến thể của Transformer.
- Xây dựng mô-đun tạo mẫu bổ sung sử dụng mạng sinh đối kháng DCGAN.
- Đánh giá độ chính xác bằng bộ dữ liệu thử nghiệm CIC-Evasive-PDFMal2022 và bộ dữ liệu tạo từ GAN.
- Giải thích quyết định của mô hình bằng những nền tảng XAI(SHAP, LIME).

Nội dung và Phương pháp

Nội dung 1: Mô-đun phát hiện tệp PDF độc hại với Transformer

- Khảo sát và tổng hợp các công trình nghiên cứu liên quan.
- Nghiên cứu cấu trúc tập tin PDF và cách rút trích các thông tin từ tập tin PDF qua thư viện PyMuPDF library.
- 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.
- Thử nghiệm các cách thức vector hóa khác nhau, đánh giá về thời gian, bộ nhớ cần thiết cho từng mô hình.

Nội dung và Phương pháp

Nội dung 2: Mô-đun sinh dữ liệu với Mạng sinh đối kháng GAN

- 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 và Phương pháp

Nội dung 3: Đánh giá, giải thích mô hình với Mạng khả diễn XAI

- Đánh giá hiệu quả mô hình với Accuracy, F1-Score, Recall, Precision.
- So sánh hiệu quả giữa bộ dữ liệu gốc, bộ dữ liệu được bổ sung bởi DCGAN.
- So sánh hiệu quả giữa các mô hình xử lý ngôn ngữ tự nhiên.
- Sử dụng các nền tảng XAI như LIME, SHAP nhằm mô tả các yếu tố đưa ra quyết định của mô hình, phân tích các đặc trưng quan trọng của tệp PDF độc hại.

Kết quả dự kiến

- 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 thành công các nền tảng học máy diễn giải XAI với mô hình Transformer nhằm giải thích các yếu tố/đặc trưng của tệp PDF độc hại đối với các quyết định của mô hình.

Tài liệu tham khảo

- 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.
- J. Hatwell, M. M. Gaber, and R. M. A. Azad, "CHIRPS: Explaining random forest classification," *Artificial Intelligence Review,* vol. 53, pp. 5747-5788, 2020.
- 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.
- I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access,* vol. 10, pp. 99129-99149, 2022.
- M. Issakhani, P. Victor, A. Tekeoglu, and A. H. Lashkari, "PDF Malware Detection based on Stacking Learning," in *ICISSP*, 2022, pp. 562-570.

Tài liệu tham khảo

- 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.
- Artifex. (2024, May 10). *PyMuPDF 1.24.3 documentation*. Available: https://pymupdf.readthedocs.io/en/latest/
- 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.
- S. Lundberg. (2023, October, 10) *SHapley Additive exPlanations*. Available: https://shap.readthedocs.io/en/latest/