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

Trường ĐH Công Nghệ Thông Tin-Đại học Quốc gia

Mục tiêu

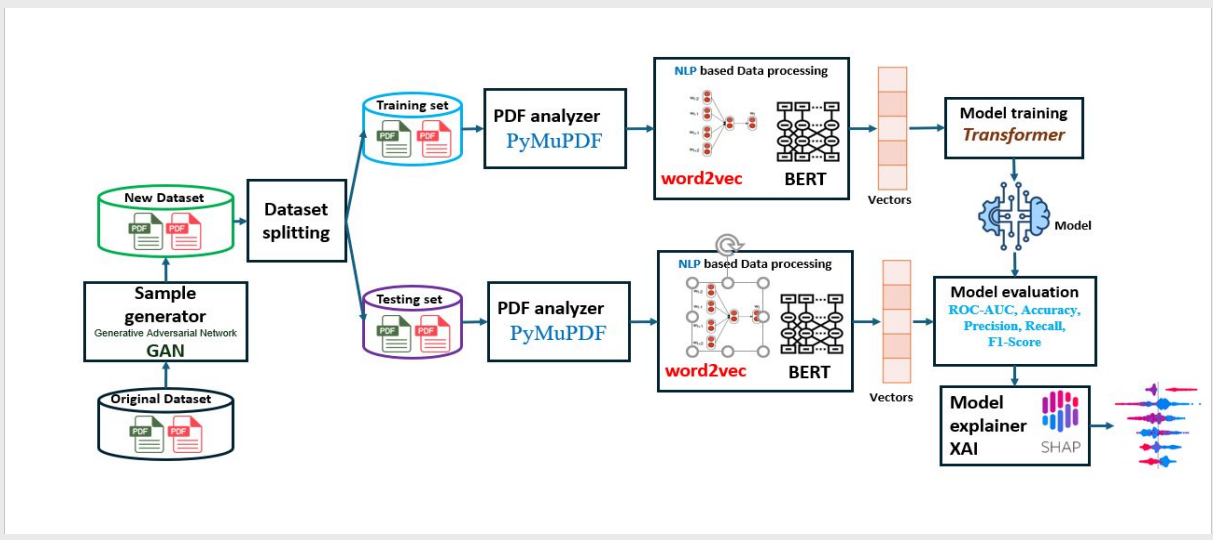
Nghiên cứu này đề xuất 1 phương pháp phát hiện tập tin PDF độc hại đáp ứng các mục tiêu sau đây:

- 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.

Lý do chọn đề tài?

- Các cuộc tấn công mã độc sử dụng tập tin PDF ngày càng phổ biến, chiếm 66.65% các cuộc tấn công thông qua email theo nghiên cứu của Palo Alto Networks vào năm 2023.
- Tập tin PDF có thể chứa các loại thông tin dạng chuỗi trong các script đính kèm, gây khó khăn cho các mô hình CNN, Random Forest, XGBoost thông thường. Các mô hình xử lý ngôn ngữ tự nhiên Transformer có tiềm năng xử lý hiệu quả đối với loại đặc trưng này.
- Các biến thể độc hại ngày càng tinh vi, xuất hiện mới hằng ngày, dẫn đến nhu cầu về nguồn dữ liệu huấn luyện.
- Việc giải thích các yếu tố ra quyết định 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. Tuy nhiên các nghiên cứu hiện nay chưa thực sự chú tâm đến vấn đề này.

Tổng quan



Mô tả

1. GAN

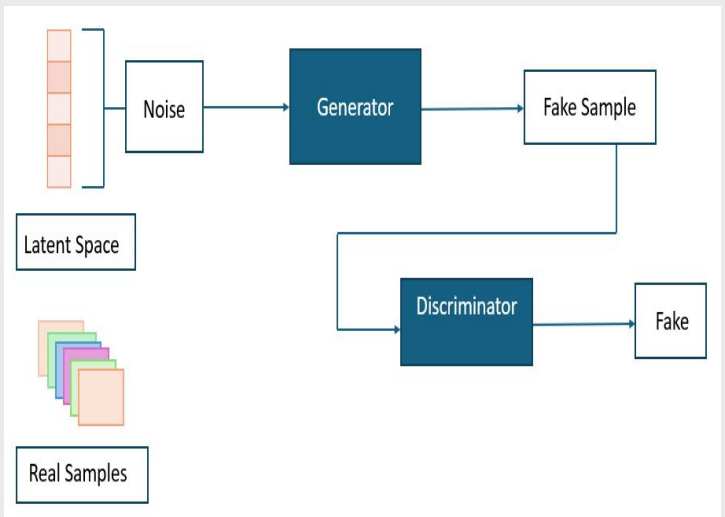
- Mạng sinh đối kháng (GAN) tạo ra các tập tin độc hại bằng cách thêm vào các nhiễu khiến các mô hình ML/DL nhầm lẫn là tập tin vô hại.
- Mô hình GAN gồm 2 mạng chạy đối lập nhau, cố gắng làm tốt hơn mạng hiện tại, gọi là bộ sinh và bộ phân biệt. Bộ sinh tìm hiểu cách thức dữ liệu được hình thành, sau đó giả định cách thức tạo ra dữ liệu. Bộ phân biệt sẽ phân loại dựa trên đầu vào của dữ liệu.

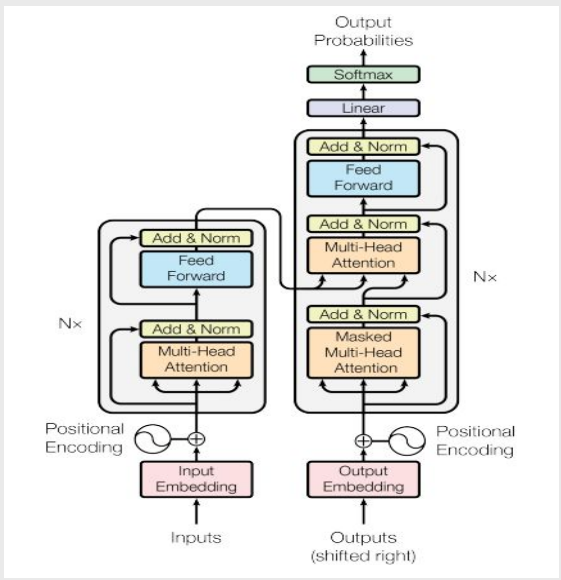
2. Transformer

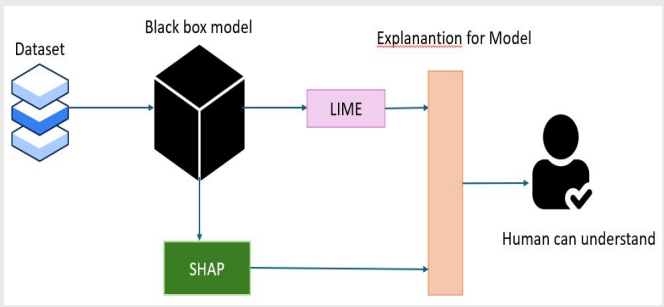
- Nghiên cứu so sánh hiệu quả giữa các mô hình BERT, DistilBERT, RoBERT, XLNet trong việc phát hiện tập tin PDF độc hại. Các đặc trưng được rút trích từ tập tin PDF qua thư viện PyMuPDF.
- Các mô hình Transformer gồm thành hai phần chính là Encoder và Decoder.
- Encoder xử lý dữ liệu đầu vào, nén dữ liệu vào vùng nhớ nhằm Decoder có thể sử dụng sau đó. Decoder nhận đầu vào từ đầu ra của Encoder kết hợp với một chuỗi đầu vào khác (Target) để tạo ra chuỗi đầu ra cuối cùng.
- Mỗi Encoder và Decoder bao gồm nhiều lớp, mỗi lớp chứa các self-attention và feed-forward neural networks.

3. XAI

- Các nền tảng khả diễn giải XAI (Explainable Artificial Intelligence) là có khả năng giải thích logic và cách thức ra quyết định của mô hình.
- Nghiên cứu sử dụng 2 mô hình XAI phổ biến là LIME (Local Interpretable Model-agnostic Explanations) và SHAP (SHapley Additive exPlanations).
- LIME tập trung vào việc diễn giải dự đoán bằng cách xây dựng một mô hình diễn giải cục bộ tại các điểm dữ liệu cụ thể.
- SHAP sử dụng lý thuyết trò chơi để phân tích đóng góp của từng đặc trưng vào dự đoán tổng thể của mô hình.







NII

Nguyễn Cẩm Hồng Phước – Trường ĐH CNTT-Đại học Quốc gia

TEL : 0845646282 Email : phuocnch.18@grad.uit.edu.vn