Phát hiện tập tin PDF độc hại sử dụng mô hình Transformer và trí tuệ nhân tạo khả diễn giải

Nguyễn Cẩm Hồng Phước[1,2]

[1] Trường Đại học Công nghệ thông tin, Thành phố Hồ Chí Minh, Việt Nam

[2] Trường Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

Mục tiêu

Nghiên cứu này đề xuất hệ thống cho phép phát hiện tập tin PDF độc hại sử dụng mô hình Transformer và trí tuệ nhân tạo khả diễn giải (XAI):

- Xây dựng phương pháp phát sinh tập tin PDF bằng mạng sinh đối kháng để tăng số lượng mẫu thử.
- Xây dựng phương pháp vector hóa các đặc trưng bằng các mô hình xử lý ngôn ngữ tự nhiên như word2vec, BERT, DistilBERT, RoBERT, XLNet
- Huấn luyện và tái huấn luyện mô hình phát hiện tập tin PDF độc hại bằng các biến thể của Transformer trên cả tập dataset gốc và dataset phát sinh từ GAN
- Đánh giá mô hình bằng các độ đo Accuracy, Precision, Recall, F1-Score. Phân tích đóng góp của từng loại đặc trưng bằng XAI.

Lý do chọn đề tài?

- Các nghiên cứu hiện tại chỉ dùng một trong các mô hình xử lý ngôn ngữ tự nhiên để rút trích đặc trưng → cần đánh giá nhiều mô hình NLP cho bài toán phát hiện tập tin PDF độc hại.
- Bộ dataset hiện tại làm các mô hình có khả năng bị overfit → cần bổ sung thêm số lượng mẫu thử để tránh overfit mô hình. → cần dùng kỹ thuật hiện đại như GAN.
- Các nghiên cứu hiện tại chưa tập trung nhiều vào việc giải thích sự đóng góp của các đặc trưng lên các mô hình → Dùng XAI để giải thích.

Tổng quan

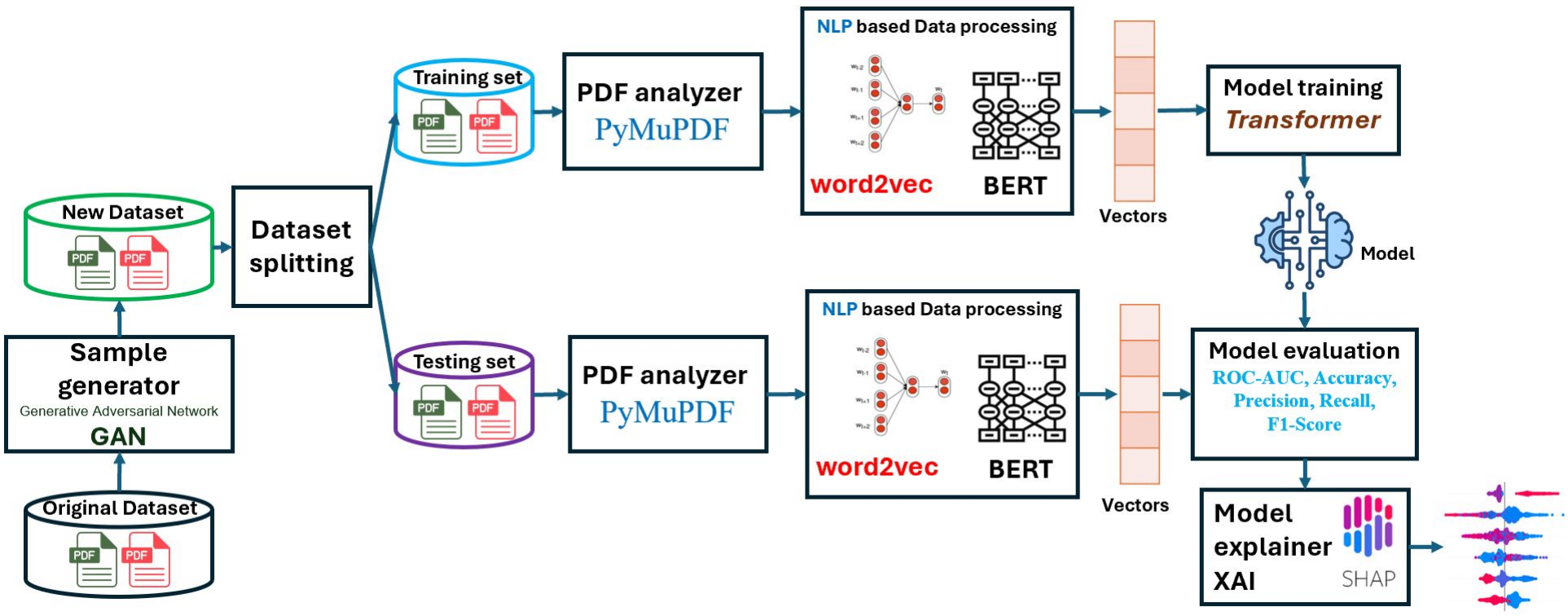

Figure 1. Sơ đồ hệ thống đề xuất

Mô tả

1. Dataset splitting

- Chia bộ dữ liệu thử nghiệm thành các tập Training set, Test set để huấn luyện mô hình và đánh giá mô hình.
- Sử dụng hướng tiếp cận 80% cho Train set và 20% cho Test set.

2. Sample generator

- Sử dụng bộ dữ liệu gốc với mô hình GAN như DCGAN để phát sinh các câu truy vấn làm mẫu thử (adversarial samples).
- Tạo bộ dữ liệu thử nghiệm mới với số lượng câu truy vấn và loại câu truy vấn nhiều hơn.

3. PDF analyzer

- Sử dụng PyMuPDF để phân tích tập tin PDF.
- Các thông tin từ tập tin PDF được rút trích từ PyMuPDF sẽ dùng để tạo vector đặc trưng.

4. NLP based Data processing

- Sử dụng cả Word2Vec, BERT, RoBERT, XLNET, DistilBERT để rút trích vector đặc trưng.
- So sánh các vấn đề về độ chính xác, tốc độ rút trích vector đặc trưng và chi phí bộ nhớ cho từng mô hình.

4. Model training

- Huấn luyện mô hình phát hiện tập tin PDF độc hại bằng các biến thể của Transformer.
- Thống kê các cách thức sử dụng các giá trị khác nhau của hyperparameters của các mô hình để phục vụ cho việc so sánh đánh giá và lựa chọn tham số tối ưu.

5. Model evaluation

- Đánh giá các mô hình được huấn luyện từ mô-đun Model training bằng các độ đo: Accuracy, Precision, Recall, F1-Score
- Phát sinh các biểu đồ, đồ thị để minh họa kết quả thử nghiệm thông qua Confusion matrix, biểu đồ độ chính xác,...

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total number of samples}} \qquad \text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \qquad \textit{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. Model explainer

- Áp dụng XAI như SHAP để giả thích sự đóng góp của các đặc trưng lên từng mô hình để lựa chọn đặc trưng phù hợp cho bài toán phát hiện tập tin PDF độc hại.

Nguyễn Cẩm Hồng Phước– Trường Đại học Công nghệ thông tin, ĐHQG TP.HCM, Việt Nam
TEL : ******* Email : phuocnch.18@grad.uit.edu.vn**