Phát hiện tấn công SQL injection sử dụng mô hình xử lý ngôn ngữ tự nhiên và mạng sinh đối kháng

Đồng Thị Ngọc Trâm

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

What ?

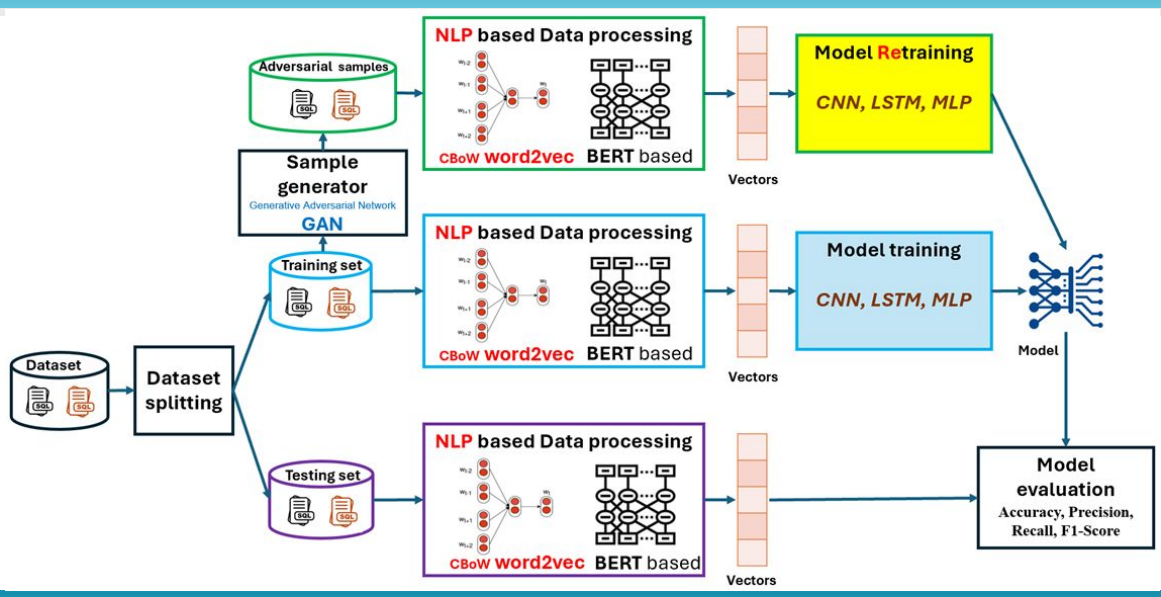
Nghiên cứu này đề xuất phát hiện tấn công SQL injection sử dụng mô hình xử lý ngôn ngữ tự nhiên và mạng sinh đối kháng (GAN):

- Xây dựng phương pháp phát sinh câu truy vấn SQL bằng mạng sinh đối kháng
- Xây dựng phương pháp rút trích vector đặc trưng bằng các mô hình xử lý ngôn ngữ tự nhiên
- Huấn luyện và tái huấn luyện mô hình phát hiện tấn công SQL injection trên cả tập dataset gốc và dataset phát sinh từ GAN
- Đánh giá mô hình bằng các độ đo Accuracy, Precision, Recall, F1-Score

Why ?

- 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 SQL injection
- Bộ dataset hiện tại làm các mô hình có khả năng bị overfit -> cần bổ sung thêm số lượng mẫu thử cũng như đa dạng các câu truy vấn độc hại để tránh overfit mô hình -> cần dùng kỹ thuật hiện đại như GAN.

Overview



Description

1. Dataset splitting

- The 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. NLP based Data processing

- Sử dụng cả Word2Vec, BERT, DistilBERT để rút trích vector đặc trưng.
- Dùng BERT vì có thể nó hiệu quả hơn trong việc hiểu ngữ nghĩa câu truy vấn SQL so với Word2Vec; dung DistilBERT vì sẽ nhanh hơn so với BERT

4. Model training

- Huấn luyện mô hình phát hiện SQL injection bằng các mô hình học sâu như: CNN, LSTM, MLP.
- 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 retraining

- Tái huấn luyện lại các mô hình bằng các mẫu thử phát sinh từ GAN (Adversarial samples)

6. Model evaluation

- Đánh giá các mô hình được huấn luyện từ mô-đun Model training và mô-đun Model Retraining 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,…

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

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

NII

Đồng Thị Ngọc Trâm – Trường Đại học Công nghệ thông tin, Thành phố Hồ Chí Minh, Việt Nam

TEL : 0975392234 Email : tramdtn.18@ms.uit.edu.vn