


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): <https://youtu.be/D9p1ekTGkgc>
(ví dụ: <https://www.youtube.com/watch?v=AWq7uw-36Ng>)
- Link slides (dạng .pdf đặt trên Github):
<https://github.com/TramDong/CS2205.MAR2024/>
(ví dụ: <https://github.com/mynameuit/CS2205.APR2023/TenDeTai.pdf>)
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">• Họ và Tên: Đồng Thị Ngọc Trâm• MSSV: 230201057 	<ul style="list-style-type: none">• Lớp: CS2205.MAR2024• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 3• Link Github:
--	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN TẤN CÔNG SQL INJECTION SỬ DỤNG MÔ HÌNH XỬ LÝ NGÔN
NGỮ TỰ NHIÊN VÀ MẠNG SINH ĐỐI KHÁNG

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

SQL INJECTION ATTACK DETECTION USING NATURAL LANGUAGE
PROCESSING AND GENERATIVE ADVERSARIAL NETWORK

TÓM TẮT (Tối đa 400 từ)

Tấn công SQL injection luôn đứng vị trí số một trong nhóm mười nguy cơ bảo mật phổ biến nhất đối với các ứng dụng web theo đánh giá của tổ chức OWASP (Open Web Application Security Project). Có nhiều nghiên cứu trong lĩnh vực phát hiện tấn công SQL injection. Một số dùng kỹ thuật phát hiện từ khóa, một số dùng mô hình học máy. Tuy nhiên, cần thiết phải áp dụng các phương pháp tiên tiến, hiện đại hơn để giải quyết các hạn chế của nghiên cứu hiện tại, tăng tính chính xác cũng như khả năng ứng dụng trong thực tế. Do các câu truy vấn SQL và các câu truy vấn HTTP (HTTP request) có dạng các chuỗi ký tự. Do đó, sử dụng các mô hình ngôn ngữ cho việc phân tích các chuỗi ký tự này được kỳ vọng sẽ mang lại nhiều hiệu quả trong việc phát hiện tấn công SQL injection. Nghiên cứu này đề xuất hệ thống phát hiện tấn công SQL injection sử dụng các mô hình xử lý ngôn ngữ tự nhiên và mạng sinh đối kháng. Cụ thể, nghiên cứu sử dụng các mô hình xử lý ngôn ngữ tự nhiên như Word2Vec, BERT, DistilBERT để vector hóa các câu truy vấn SQL thành các vector đặc trưng. Các vector đặc trưng này được dùng để huấn luyện các mô hình học sâu như CNN (Convolutional Neural Network), LSTM (Long Short Term Memory), MLP (Multilayer Perceptron) để xây dựng mô hình phát hiện tấn công SQL injection. Các nghiên cứu hiện tại chủ yếu sử dụng bộ dữ liệu thử nghiệm từ Kaggle với 19.537 câu truy vấn. Nghiên cứu này đề xuất sử dụng mạng sinh đối kháng (GAN) để phát sinh câu truy vấn SQL độc hại và lành tính để bổ sung mẫu thử cho bộ dữ liệu thử nghiệm

này nhằm huấn luyện lại các mô hình nhằm cải thiện độ chính xác khi đánh giá các câu truy vấn SQL ngoài thực tế. Thay vì dùng GAN để tạo các vector đặc trưng tùy biến từ các vector đặc trưng gốc, nghiên cứu này dùng GAN để tạo các câu truy vấn SQL tùy biến từ các câu SQL gốc. Đây là một trong những điểm nổi bật của nghiên cứu này.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Theo đánh giá của dự án OWASP [1], tấn công SQL injection luôn đứng vị trí đầu tiên trong nhóm các nguy cơ bảo mật trong ứng dụng web. Tấn công SQL injection diễn ra khi kẻ tấn công chèn thêm các thành phần của câu truy vấn SQL để tạo các câu truy vấn SQL độc hại để đánh cắp dữ liệu hoặc gây hư hại dữ liệu của các ứng dụng web.

Có nhiều nghiên cứu liên quan đến việc xây dựng mô hình phát hiện tấn công SQL injection. Lakhani và đồng sự [2] đề xuất hệ thống phát hiện SQL injection bằng cách dùng BERT để tiền xử lý dữ liệu và dùng các mô hình học máy SVM, Random Forest. Tuy nhiên, BERT [3] là một mô hình tốn nhiều thời gian trong quá trình huấn luyện. Nghiên cứu này có thể được mở rộng bằng cách dùng các kỹ thuật tiền xử lý dữ liệu khác như DistilBERT [4] và Word2Vec [5] để đánh giá sự khác biệt giữa các phương pháp này. Trong khi đó, Natanajan và đồng sự [6] sử dụng Logistic Regression, Naive Bayes, Random Forest và Convolutional Neural Network (CNN) để phát hiện tấn công SQL injection. Kết quả thử nghiệm của họ cho thấy CNN có độ chính xác cao nhất với 99.29%. Tuy nhiên, cần thiết phải thực hiện việc thử nghiệm sử dụng các cách thức rút trích vector đặc trưng khác nhau thông qua BERT, DistilBERT và Word2Vec. Đồng thời sử dụng các mô hình học sâu khác như LSTM, MLP.

Trong một nghiên cứu khác, Gogoi và đồng sự [7] cũng đề xuất hệ thống phát hiện SQL injection dựa vào NLP. Trong nghiên cứu này, họ dùng các mô hình học máy và recall và f1-score đạt 99.9%. Cũng giống như các nghiên cứu khác, nghiên cứu này cần được thử nghiệm với nhiều mô hình rút trích vector đặc trưng khác nhau. Fang và cộng sự [8] sử dụng word vector cùng với LSTM để phát hiện tấn công SQL injection. Kết quả thử nghiệm của họ cho thấy độ chính xác đạt 98.6%. Tuy nhiên, cần phải sử dụng các phương pháp rút trích vector đặc trưng khác như BERT hay DistilBERT và so sánh với các mô hình học sâu khác.

Dữ liệu huấn luyện là một trong những yếu tố quan trọng trong việc xây dựng các mô

hình học sâu. Trong bài toán phát hiện tấn công SQL injection, bộ dữ liệu thử nghiệm từ Kaggle cần được bổ sung các mẫu thử không những tăng về số lượng mà còn đa dạng các kịch bản tấn công. Việc ứng dụng mạng sinh đối kháng (GAN) [9] để tạo các mẫu thử cho tập dữ liệu này là cần thiết và đó cũng là một đóng góp dự kiến của nghiên cứu này.

Nghiên cứu này có các đóng góp chính dự kiến bao gồm: (1) Đề xuất mô-đun rút trích đặc trưng bằng cách dùng mô hình xử lý ngôn ngữ tự nhiên như Word2Vec, BERT, DistilBERT, đồng thời đánh giá, so sánh chúng một cách chi tiết. (2) Đề xuất mô-đun phát sinh câu truy vấn SQL bằng mạng sinh đối kháng (GAN) để gia tăng số lượng mẫu thử cho dataset, đồng thời tái huấn luyện mô hình bằng bộ dữ liệu mới nhằm tăng độ chính xác khi phân tích các câu truy vấn SQL ngoài thực tế. (3) Huấn luyện mô hình phát hiện tấn công SQL injection sử dụng mô hình học sâu như CNN, LSTM, MLP bằng cả bộ dữ liệu gốc và bộ dữ liệu thử nghiệm phát sinh bằng GAN trong nghiên cứu .

MỤC TIÊU

- Xây dựng được mô-đun rút trích đặc trưng bằng mô hình xử lý ngôn ngữ tự nhiên như Word2Vec, BERT, DistilBERT để rút trích vector đặc trưng từ các câu truy vấn SQL hoặc HTTP request. Đánh giá, so sánh các mô hình để chọn mô hình phù hợp cho từng loại môi trường triển khai.
- Xây dựng được mô-đun tạo các câu truy vấn SQL làm mẫu thử bổ sung cho dataset tấn công SQL injection hiện tại bằng cách sử dụng mạng sinh đối kháng như DCGAN.
- Xây dựng được mô hình phát hiện tấn công SQL injection bằng các mô hình học sâu như CNN, LSTM, MLP trên bộ dữ liệu thử nghiệm gốc từ Kaggle và bộ dữ liệu được tạo từ GAN trong nghiên cứu .

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung nghiên cứu 1: Xây dựng mô-đun rút trích đặc trưng bằng mô hình xử lý ngôn ngữ tự nhiên như Word2Vec, BERT, DistilBERT để rút trích vector đặc trưng từ các câu truy vấn SQL hoặc HTTP request. Đánh giá, so sánh các mô hình để chọn mô hình phù hợp cho từng loại môi trường triển khai.

Phương pháp nghiên cứu:

- Tìm hiểu cấu trúc các lệnh SQL, các kỹ thuật tấn công SQL injection như Boolean based, Union based, error based,... để phục vụ cho việc tiền xử lý trước khi thực hiện vector hóa đặc trưng.
- Khảo sát và tổng hợp các công trình nghiên cứu liên quan (literature review) về các phương pháp phát hiện tấn công SQL injection để xác định các khoảng trống nghiên cứu (research gaps).
- Nghiên cứu cách thức vector hóa các đặc trưng bằng các mô hình xử lý ngôn ngữ tự nhiên như word2vec, BERT, DistilBERT.
- Thử nghiệm các cách thức vector hóa khác nhau, đánh giá về thời gian, bộ nhớ cần thiết cho từng mô hình.

Nội dung nghiên cứu 2: Xây dựng mô-đun tạo các câu truy vấn SQL làm mẫu thử bổ sung cho dataset hiện tại bằng cách sử dụng mạng sinh đối kháng như DCGAN.

Phương pháp nghiên cứu:

- 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 câu truy vấn SQL tùy chỉnh từ GAN.
- Kiểm thử chất lượng mẫu thử được tạo từ GAN.
- Chuẩn hóa dataset mới được tạo từ GAN để cung cấp cho các nghiên cứu liên quan trong tương lai.

Nội dung nghiên cứu 3: Xây dựng mô hình phát hiện tấn công SQL injection bằng các mô hình học sâu như CNN, LSTM, MLP trên bộ dữ liệu thử nghiệm gốc từ Kaggle và bộ dữ liệu được tạo từ GAN trong nghiên cứu .

Phương pháp nghiên cứu:

- Xây dựng các mô hình nhận diện tấn công SQL injection dựa vào các mô hình học sâu như CNN, LSTM, MLP bằng cách huấn luyện từ bộ dữ liệu gốc và bộ dữ liệu thử nghiệm phát sinh từ GAN.
- Tinh chỉnh các siêu tham số của các mô hình để xác định bộ siêu tham số tối ưu

- Sử dụng các độ đo như Accuracy, F1-Score, Recall, Precision để đánh giá mô hình được huấn luyện từ dataset gốc và dataset được tạo từ GAN.

Kế hoạch thực hiện:

Mục tiêu đặt ra là hoàn thành nghiên cứu trong khoảng 12 tháng. Kế hoạch thực hiện các nội dung (ND) chính như sơ đồ Gantt sau.

	Tháng											
Công việc	1	2	3	4	5	6	7	8	9	10	11	12
ND1												
ND2												
ND3												

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

Các kết quả mong đợi của luận văn bao gồm:

- Xây dựng được hệ thống phát hiện tấn công SQL injection sử dụng các mô hình xử lý ngôn ngữ tự nhiên khác nhau và có khả năng tái huấn luyện dựa vào mẫu thử phát sinh từ mạng sinh đối kháng. Độ chính xác mô hình tăng khi tái huấn luyện với dataset mới.
- Xây dựng được mô-đun phát sinh các câu truy vấn SQL làm mẫu thử bằng cách dùng mạng sinh đối kháng.

TÀI LIỆU THAM KHẢO (Định dạng DBLP)

- [1] OWASP. (2024, May 01). *OWASP Top Ten*. Available: <https://owasp.org/www-project-top-ten/>
- [2] S. Lakhani, A. Yadav, and V. Singh, "Detecting SQL Injection Attack using Natural Language Processing," in *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2022, pp. 1-5.
- [3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bidirectional encoder representations from transformers," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1, pp. 4171-4186.
- [4] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *Journal of big Data*, vol. 11, no. 1, p. 25, 2024.

- [5] Q. Jiao and S. Zhang, "A brief survey of word embedding and its recent development," in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2021, vol. 5, pp. 1697-1701: IEEE.
- [6] Y. Natarajan, B. Karthikeyan, G. Wadhwa, S. A. Srinivasan, and A. S. P. Akilesh, "A Deep Learning Based Natural Language Processing Approach for Detecting SQL Injection Attack," Cham, 2023, pp. 396-406: Springer Nature Switzerland.
- [7] B. Gogoi, T. Ahmed, and A. Dutta, "Defending against SQL Injection Attacks in Web Applications using Machine Learning and Natural Language Processing," in *2021 IEEE 18th India Council International Conference (INDICON)*, 2021, pp. 1-6.
- [8] Y. Fang, J. Peng, L. Liu, and C. Huang, "WOVSQLI: Detection of SQL Injection Behaviors Using Word Vector and LSTM," presented at the Proceedings of the 2nd International Conference on Cryptography, Security and Privacy, Guiyang, China, 2018. Available: <https://doi.org/10.1145/3199478.3199503>
- [9] 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.