THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube:
  https://www.youtube.com/watch?v=YnO4q4YSuDs
- Link slides:
  https://github.com/mynameuit/CS2205.xxx/TenDeTai.pdf
  https://github.com/Le-Ton-Nhan/CS2205.FEB2025/XÂY DỰNG GIẢI PHÁP PHÁT HIỆN LIÊN KẾT LỪA ĐẢO SỬ DỤNG MÁY HỌC.pdf

- Họ và Tên: Lê Tôn Nhân
- MSSV: 240202025

- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 15
- Link Github:
  https://github.com/Le-Ton-Nhan/CS2205.FEB2025

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

XÂY DỰNG GIẢI PHÁP PHÁT HIỆN LIÊN KẾT LỪA ĐẢO SỬ DỤNG MÁY HỌC

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

BUILDING A MACHINE LEARNING-BASED SOLUTION FOR DETECTING PHISHING LINK

TÓM TẮT *(Tối đa 400 từ)*

Trong bối cảnh tấn công lừa đảo ngày càng tinh vi và đa dạng, việc phát hiện sớm các liên kết lừa đảo đóng vai trò then chốt trong bảo vệ người dùng và hệ thống. Để đối phó với vấn đề này, chúng tôi đề xuất một hệ thống phát hiện liên kết lừa đảo bằng cách kết hợp hai hướng tiếp cận: (1) phân tích thông tin tĩnh của liên kết bằng các thuật toán học máy có giám sát (Random Forest), và (2) phân tích thông tin giao diện dựa trên logo trang web sử dụng mô hình học sâu (Object Detection kết hợp Siamese Network). Hệ thống được triển khai dưới dạng ứng dụng web, cho phép người dùng gửi các liên kết nghi vấn và nhận kết quả phân tích nhanh chóng. Phương pháp này giúp tăng độ chính xác, giảm phụ thuộc vào dữ liệu gán nhãn và đối phó tốt hơn với các kỹ thuật né tránh của kẻ tấn công.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Tấn công lừa đảo đã phát triển một cách nhanh chóng, đến mức hiện nay, nó được cung cấp như một dịch vụ [1]. Trước tình hình đó, các nhà nghiên cứu đã đề xuất nhiều phương pháp phát hiện khác nhau, trong đó nổi bật là hai hướng chính: dựa trên thông tin tĩnh và dựa trên thông tin giao diện.

Phát hiện liên kết lừa đảo dựa trên thông tin tĩnh là một phương pháp phổ biến, sử dụng mô hình học máy để phân tích đặc trưng như URL hoặc mã HTML. Tuy nhiên, cách tiếp cận này phụ thuộc nhiều vào dữ liệu gán nhãn và dễ bị kẻ tấn công qua mặt bằng các kỹ thuật né tránh, làm giảm hiệu quả trước các mối đe dọa mới.

Phương pháp phát hiện dựa trên thông tin giao diện so sánh hình ảnh trang web (thường là logo) với cơ sở dữ liệu thương hiệu chính thống [2]. Nếu logo giống nhau nhưng tên miền khác, hệ thống có thể xác định đó là trang lừa đảo. Tuy nhiên, cách tiếp cận này cũng gặp khó khăn khi giao diện web thay đổi liên tục và cần cập nhật dữ liệu thường xuyên.

Một hướng tiếp cận mới từ năm 2021 là sử dụng hệ thống học sâu kết hợp (hybrid deep learning), gồm mô hình nhận diện logo và mô hình Siamese để xác định thương hiệu mà không cần dữ liệu lừa đảo, giúp giảm lệch dữ liệu và tiết kiệm chi phí. Ngoài ra, kỹ thuật gradient masking được tích hợp nhằm tăng cường khả năng chống né tránh tấn công [3].

Với đề tài này, chúng tôi phát triển hệ thống phát hiện liên kết lừa đảo kết hợp cả thông tin tĩnh và giao diện, nhằm giảm chi phí và nâng cao hiệu quả, độ chính xác, từ đó mang lại một giải pháp bảo vệ người dùng trực tuyến đáng tin cậy hơn.

Input: đoạn văn bản hoặc email, tin nhắn SMS chứa liên kết muốn kiểm tra.

Output: kết quả dự đoán liên kết có phải là lừa đảo hay không, cùng các đặc trưng trích xuất từ liên kết được cung cấp.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

Cung cấp một hệ thống phát hiện hiệu quả và chính xác để xác định liên kết lừa đảo, giúp người dùng tránh tiếp cận và rơi vào các cuộc tấn công lừa đảo trực tuyến. Hệ thống được thiết kế để cung cấp kết quả phát hiện lừa đảo chính xác và tin cậy. Từ việc phân tích đoạn văn bản và các yếu tố liên quan, hệ thống sẽ đưa ra đánh giá về tính xác thực của liên kết được kiểm tra.

Các module trong hệ thống được thiết kế tường minh và mở rộng, nhằm hỗ trợ cho các nghiên cứu tiếp theo và phát triển công nghệ phát hiện lừa đảo. Điều này đảm bảo tính linh hoạt và khả năng tiến bộ của hệ thống, đồng thời giúp nâng cao hiệu quả và chính xác trong phát hiện liên kết lừa đảo.

Nâng cao chất lượng bộ dữ liệu, đảm bảo tính đáng tin cậy và đa dạng, tạo cơ sở cho các nghiên cứu và phát triển tiếp theo trong lĩnh vực phát hiện liên kết lừa đảo. Sự cải

thiện bộ dữ liệu không chỉ giúp đánh giá chính xác hơn hiệu suất của hệ thống, mà còn tạo ra cơ sở cho các nghiên cứu và phát triển tiếp theo. Nhờ đó, những nghiên cứu sau này có thể dễ dàng sử dụng và mở rộng bộ dữ liệu, đồng thời nâng cao khả năng áp dụng và hiệu quả của các giải pháp phát hiện liên kết lừa đảo.

NỘI DUNG VÀ PHƯƠNG PHÁP

Hệ thống chúng tôi đề xuất cung cấp quy trình phát hiện liên kết lừa đảo toàn diện. Trước tiên, hệ thống chụp màn hình liên kết trong môi trường sandbox và trích xuất nhiều đặc trưng khác nhau của liên kết lừa đảo.

Hệ thống gồm hai nhánh: (1) Nhánh giao diện sử dụng mô hình học sâu kết hợp (Hybrid deep learning) để phát hiện dựa trên hình ảnh; (2) Nhánh tĩnh sử dụng mô hình học máy có giám sát, như Random Forest, để phân tích đặc trưng kỹ thuật. Kết hợp kết quả từ hai nhánh giúp hệ thống đánh giá chính xác độ tin cậy của liên kết và bảo vệ người dùng khỏi các rủi ro lừa đảo trực tuyến.

Nhánh phát hiện lừa đảo từ thông tin giao diện (1) tập trung vào việc nhận dạng logo trên trang web. Nếu logo của một thương hiệu xuất hiện trên trang không thuộc sở hữu của thương hiệu đó, trang có thể bị coi là lừa đảo.

Quá trình gồm hai bước: (i) Phát hiện logo và ô nhập liệu bằng mô hình học sâu (sử dụng ResNet50, RPN và Fast-RCNN) [4][5][6], và (ii) So sánh logo phát hiện được với danh sách logo thương hiệu mục tiêu bằng mô hình Siamese để xác định tên miền dự kiến [7][8]. Nếu tên miền thực tế không khớp với tên miền dự kiến, liên kết bị đánh giá là lừa đảo.

Hệ thống sau đó kết hợp logo, ô nhập liệu và kết quả phát hiện lừa đảo để tạo giải thích trực quan, giúp người dùng dễ hiểu và cảnh giác trước các trang web giả mạo [9][10].

Trong nhánh phát hiện lừa đảo từ thông tin tĩnh (2), hệ thống trích xuất nhiều đặc trưng khác nhau từ liên kết đầu vào, chia thành ba nhóm chính:

Lexical: Đặc trưng từ chuỗi liên kết như độ dài URL, số lượng chữ số, tham số,... phản ánh cấu trúc đường dẫn.

External service: Thông tin từ máy chủ như quốc gia đăng ký, tên miền, cổng mở, thời gian tồn tại,... giúp nhận diện nguồn gốc liên kết.

Content-based: Phân tích mã HTML để phát hiện các yếu tố đáng ngờ như thẻ script, tệp nhúng, đối tượng ẩn,...

Các đặc trưng này là đầu vào cho mô hình học máy có giám sát. Trong đó, Random Forest là lựa chọn phù hợp nhờ khả năng xây dựng nhiều cây quyết định và tổng hợp kết quả để phân loại liên kết là an toàn hay lừa đảo. Mô hình được huấn luyện trên tập dữ liệu đã gán nhãn do nhóm tự thu thập.

Cuối cùng, kết quả từ hai nhánh được kết hợp và hiển thị trên giao diện web trực quan, giúp người dùng dễ dàng truy cập và sử dụng hệ thống một cách tiện lợi qua trình duyệt.

KẾT QUẢ MONG ĐỢI

Hệ thống được phát triển dưới dạng ứng dụng web với 4 module chính:

Xử lý dữ liệu đầu vào: Trích xuất liên kết từ văn bản người dùng nhập và thu thập các thông tin liên quan như tên file, host. Từ đó, xác định đặc trưng và tạo cơ sở dữ liệu ban đầu cho phân tích.

Phát hiện từ thông tin tĩnh: Sử dụng các đặc trưng trích xuất từ liên kết làm đầu vào cho mô hình học máy (Random Forest) để phân loại liên kết là lừa đảo hay không.

Phát hiện từ giao diện: Dùng Selenium để chụp ảnh trang web và đưa vào mô hình học sâu đã huấn luyện, dựa trên logo và ô nhập liệu để nhận diện trang lừa đảo.

Tổng hợp kết quả: Kết hợp kết quả từ hai nhánh bằng voting hoặc xác suất, sau đó hiển thị kết luận và thông tin chi tiết trên giao diện web.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, Gail-Joon Ahn: Sunrise to Sunset: Analyzing the End-to-end Life Cycle and Effectiveness of Phishing Attacks at Scale. USENIX Security Symposium 2020: 361-377

[2] Changbo Hu, Qun Li, Zhen Zhang, Keng-hao Chang, Ruofei Zhang: A Multimodal Fusion Framework for Brand Recognition from Product Image and Context. ICME Workshops 2020: 1-4
[3] Penghui Zhang, Adam Oest, Haehyun Cho, Zhibo Sun, RC Johnson, Brad Wardman, Shaown Sarker, Alexandros Kapravelos, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, Adam Doupé, Gail-Joon Ahn: CrawlPhish: Large-scale Analysis of Client-side Cloaking Techniques in Phishing. SP 2021: 1109-1124
[4] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. CVPR 2023: 7464-7475
[5] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun: YOLOX: Exceeding YOLO Series in 2021. CoRR abs/2107.08430 (2021)
[6] Chien-Yao Wang, I-Hau Yeh, Hong-Yuan Mark Liao: You Only Learn One Representation: Unified Network for Multiple Tasks. CoRR abs/2105.04206 (2021)
[7] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, Jian Sun: You Only Look One-Level Feature. CVPR 2021: 13039-13048
[8] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao: YOLOv4: Optimal Speed and Accuracy of Object Detection. CoRR abs/2004.10934 (2020)
[9] Yun Lin, Jun Sun, Gordon Fraser, Ziheng Xiu, Ting Liu, and Jin Song Dong. Recovering fitness gradients for interprocedural boolean flags in search-based testing. In ISSTA, pages 440–451, 2020.
[10] Diego A. Velázquez, Josep M. Gonfaus, Pau Rodríguez, F. Xavier Roca, Seiichi Ozawa, Jordi Gonzàlez: Logo Detection With No Priors. IEEE Access 9: 106998-107011 (2021)