ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN TRONG NHẬN DIỆN VÀ PHÂN TÍCH TRANG LỪA ĐẢO FACEBOOK

Hoàng Minh Thành[1], Lê Công Thành[1], Cấn Anh Minh[1]
Đào Thị Phương Anh[2,*], Lê Thị Kiều Oanh[3]
[1]Sinh viên, Trường Đại học Tài nguyên và Môi trường Hà Nội
[2]Trường Đại học Tài nguyên và Môi trường Hà Nội
[3]Trường Đại học Kinh tế Kỹ thuật Công nghiệp

Tóm tắt

Nghiên cứu này giải quyết thách thức về vấn đề lừa đảo trực tuyến đang ngày càng tinh vi trên nền tảng mạng xã hội Facebook tại Việt Nam. Mục đích của nghiên cứu là xây dựng một hệ thống nhận diện và phân tích lừa đảo không chỉ dựa trên độ chính xác kỹ thuật mà còn hướng tới tính minh bạch và khả năng giải thích (Explainable AI). Đối tượng nghiên cứu là các trang cá nhân và Fanpage công khai trên Facebook với các kịch bản lừa đảo phổ biến như cờ bạc và tuyển dụng giả mạo. Phương pháp nghiên cứu được đề xuất là một kiến trúc phân tầng tích hợp mô hình lai PhoBERT-v2 kết hợp XGBoost để xử lý đồng thời đặc trưng ngữ nghĩa tiếng Việt và metadata hành vi. Hệ thống ứng dụng kỹ thuật tăng cường truy xuất tri thức (RAG) và giao thức Model Context Protocol (MCP) để xây dựng tác nhân AI có khả năng tra cứu danh sách đen thời gian thực và cung cấp các giải thích dựa trên cơ sở pháp lý. Kết quả thực nghiệm trên tập dữ liệu thực tế cho thấy mô hình đạt độ chính xác 92,58 %, F1-Score 92,76 % và ROC-AUC 0,9509. Nghiên cứu đóng góp một giải pháp toàn diện giúp nâng cao nhận thức an toàn thông tin cho cộng đồng thông qua việc kết hợp giữa công nghệ nhận diện và giáo dục người dùng.

Từ khóa: Giao thức MCP; Kỹ thuật tăng cường truy xuất RAG; Lừa đảo Facebook; Mô hình ngôn ngữ lớn LLM; Mô hình PhoBERT; Tác nhân AI.

Abstract

Application of large language models in identifying and analyzing Facebook scam pages

This study addresses the escalating challenge of sophisticated online scams on the Facebook social media platform in Vietnam. The research aims to develop a scam detection and analysis system that prioritizes not only technical accuracy but also transparency and explainability (Explainable AI). The research subjects are public Facebook profiles and Fanpages involving common scam scenarios such as gambling and fake recruitment. The proposed methodology is a layered architecture integrating a hybrid PhoBERT-v2 and XGBoost model to simultaneously process Vietnamese semantic features and behavioral metadata. The system utilizes Retrieval-Augmented Generation (RAG) and the Model Context Protocol (MCP) to build AI agents capable of real-time blacklist lookups and providing legal-based explanations. Experimental results on real-world datasets show that the model achieves 92.58 % accuracy, 92,76 % F1-Score, and a 0.9509 ROC-AUC. The study contributes a comprehensive solution

to enhance community information security awareness by combining detection technology with user education.

Keywords: MCP; RAG; Facebook Scams; LLM; PhoBERT; AI Agent.

BBT nhận bài: 31/3/2026; Phản biện xong: 28/4/2026; Chấp nhận đăng: 12/6/2026

*Tác giả liên hệ, Email: dtpanh@hunre.edu.vn

DOI: https://doi.org/10.63064/khtnmt.2026.851

1. Giới thiệu

Trong kỷ nguyên số hóa, sự phát triển mạnh mẽ của công nghệ đã thay đổi căn bản cách con người tương tác và giao tiếp trên mạng xã hội. Tuy nhiên, các nền tảng như Facebook không chỉ là nơi kết nối mà còn trở thành môi trường tiềm ẩn rủi ro lừa đảo trực tuyến ngày càng tinh vi, gây thiệt hại kinh tế ước tính gần 20.000 tỷ đồng mỗi năm tại Việt Nam. Theo thống kê từ các cơ quan chức năng, chỉ riêng trong năm 2024, đã có hơn 125.000 website giả mạo, mạo danh các cơ quan, tổ chức được ghi nhận với hàng trăm kịch bản và chiêu trò lừa đảo khác nhau [1, 23]. Quy mô của vấn đề còn được minh chứng qua dữ liệu từ Meta: Trong nửa đầu năm 2025, đã có khoảng hơn 5,4 triệu nội dung gian lận, lừa đảo và gây hiểu lầm trên Facebook tại Việt Nam bị gỡ bỏ. Những con số này phản ánh rõ rệt nhu cầu cấp bách về các giải pháp an ninh mạng hiệu quả nhằm bảo vệ người dùng [2, 20].

Thực tế cho thấy, các phương pháp phòng thủ truyền thống dựa trên tập luật (rule-based) hoặc các mô hình học máy cơ bản thường gặp nhiều hạn chế trước các kịch bản lừa đảo hiện đại [15, 25 - 27]. Tội phạm mạng hiện nay đã chuyển dịch từ việc sử dụng mã độc kỹ thuật sang các hành vi "thao túng tâm lý" tinh vi, đánh vào nỗi sợ hãi hoặc lòng tham của nạn nhân thông qua ngôn ngữ dẫn dụ. Các hệ thống hiện tại chủ yếu tập trung vào việc kiểm tra các thực thể kỹ thuật như URL độc hại hay danh sách đen, nhưng lại thiếu khả năng hiểu sâu về ngữ nghĩa của văn bản tiếng Việt để bóc tách các lớp "ngụy trang" tâm lý. Hơn nữa, sự tách rời giữa công nghệ nhận diện và việc giáo dục người dùng tạo ra một khoảng trống lớn: người dùng thường nhận được những cảnh báo kỹ thuật khô khan mà không hiểu rõ lý do tại sao một trang thông tin lại bị đánh giá là nguy hiểm.

Sự ra đời của các mô hình ngôn ngữ lớn (LLM) mở ra hướng đi mới nhờ khả năng hiểu ngữ cảnh sâu sắc. Tuy nhiên, các nghiên cứu trong nước vẫn còn hạn chế trong việc tích hợp sâu rộng các kỹ thuật xử lý tiếng Việt chuyên sâu đi kèm với cơ chế giải thích giúp con người hiểu và tin tưởng kết quả do AI nói chung tạo ra [16, 17, 21, 22].

Bài báo này đề xuất một hệ thống nhận diện và phân tích trang lừa đảo Facebook toàn diện thông qua kiến trúc phân tầng linh hoạt. Giải pháp cốt lõi dựa trên việc tích hợp mô hình lai PhoBERT-v2 kết hợp XGBoost để nắm bắt đồng thời đặc trưng ngữ nghĩa và metadata hành vi. Điểm đổi mới sáng tạo chính là việc ứng dụng kỹ thuật tăng cường truy xuất tri thức (RAG) nhằm giảm thiểu hiện tượng "ảo giác" của AI và giao thức Model Context Protocol (MCP) nhằm thiết lập cơ chế giao tiếp chuẩn hóa với các nguồn dữ liệu ngoại vi. Mục tiêu của nghiên cứu không chỉ dừng

lại ở việc nâng cao độ chính xác kỹ thuật mà còn hướng tới xây dựng một "lá chắn tri thức", giúp người dùng nhận diện các dấu hiệu thao túng và bảo vệ mình một cách chủ động hơn trên không gian mạng.

2. Cơ sở lý thuyết và phương pháp nghiên cứu

2.1. Cơ sở lý thuyết

2.1.1. Xử lý ngôn ngữ tự nhiên (NLP) và mô hình PhoBERT

Xử lý ngôn ngữ tự nhiên hiện đại dựa trên kiến trúc Transformer cho phép máy tính hiểu và phản hồi ngôn ngữ con người một cách có ngữ cảnh thông qua cơ chế self-attention. Trong nghiên cứu này, NLP đóng vai trò là "thấu kính" để bóc tách các yếu tố phi ngôn ngữ như phân tích cảm xúc (Sentiment Analysis), nhận diện ý đồ (Intent Recognition) và xác định sắc thái (Tone of voice) nhằm phát hiện các hành vi gây áp lực tâm lý trong kịch bản lừa đảo [3].

Nghiên cứu ứng dụng PhoBERT-base-v2, mô hình ngôn ngữ tiền huấn luyện chuyên biệt cho tiếng Việt được huấn luyện trên bộ dữ liệu 120 Gb. Điểm ưu việt của PhoBERT là thực hiện tách từ (word segmentation) trước khi huấn luyện, giúp khắc phục nhược điểm hiểu sai ngữ cảnh của các mô hình đa ngôn ngữ khi xử lý đặc thù tiếng Việt [4].

2.1.2. Thuật toán phân loại XGBoost

Thuật toán XGBoost là một mô hình học máy kết hợp (ensemble learning) dựa trên cây quyết định và phương pháp Gradient Boosting, được thiết kế tối ưu cho khả năng mở rộng. Hàm mất mát tổng quát của XGBoost được định nghĩa như sau:

$$L_{xgb} = \sum_{i=1}^{N} L\left(y_i, F(x_i)\right) + \sum_{m=1}^{M} \Omega(h_m) \quad (1)$$

trong đó, $L(y_i, F(x_i))$: là hàm suy hao đo lường sự sai lệch giữa giá trị dự đoán và thực tế. Phần điều chuẩn Ω(h) đóng vai trò kiểm soát độ phức tạp của mô hình, giúp tránh hiện tượng quá khớp (overfitting):

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda |w|^2 \quad (2)$$

T đại diện cho số lượng lá của cây quyết định và *w* là trọng số tại các lá.

Với đặc tính xử lý tốt dữ liệu thưa và tốc độ huấn luyện song song, XGBoost rất phù hợp để phân loại metadata hành vi của các trang Facebook trong thời gian thực.

2.1.3. Mô hình ngôn ngữ lớn (LLMs) và tác nhân AI (AI Agent)

LLMs là các mạng nơ-ron sâu với quy mô tham số cực lớn, có khả năng suy luận logic và học trong ngữ cảnh [6, 24]. Khác với các mô hình phản hồi thụ động, tác nhân AI dựa trên LLM là một thực thể chủ động thực hiện vòng lặp nhận thức - suy luận - hành động. Cấu trúc của tác nhân gồm bốn thành phần cốt lõi: mô hình suy luận (Core LLM), mô-đun lập kế hoạch, mô-đun bộ nhớ và mô-đun hành động thực thi qua API [10]. Điều này cho phép hệ thống tự động điều phối các công cụ kiểm tra rủi ro dựa trên bối cảnh người dùng.

2.1.4. Kỹ thuật RAG và giao thức ngữ cảnh mô hình (MCP)

Để khắc phục hiện tượng "ảo giác" (hallucination) của LLM, kỹ thuật tăng cường truy xuất (RAG) được ứng dụng nhằm lấy dữ liệu thực tế từ các cơ sở dữ liệu vector (như ChromaDB) để bổ

sung ngữ cảnh cho mô hình [18]. Dữ liệu được chia nhỏ thành các đoạn (chunks) và chuyển đổi thành vector nhúng để tìm kiếm ngữ nghĩa thay vì tìm kiếm từ khóa đơn thuần. Song song đó, giao thức MCP cung cấp một lớp tiêu chuẩn hóa mã nguồn mở, hoạt động như một "cổng USB-C cho AI", giúp chuẩn hóa kết nối giữa tác nhân AI và các dịch vụ bên ngoài (như danh sách đen số điện thoại, tài khoản ngân hàng) thông qua định dạng JSON-RPC 2.0 [13]. Sự kết hợp giữa RAG và MCP đảm bảo hệ thống đưa ra các giải thích minh bạch, có bằng chứng và luôn cập nhật thời gian thực.

2.2. Phương pháp nghiên cứu

2.2.1. Cách tiếp cận và quy trình nghiên cứu

Nghiên cứu áp dụng phương pháp luận hỗn hợp (mixed methods) kết hợp giữa nghiên cứu lý thuyết và thực nghiệm. Quy trình thực hiện tuân thủ chu trình phát triển hệ thống AI an ninh mạng chuẩn mực, bao gồm: (1) Phân tích yêu cầu chức năng và phi chức năng; (2) Thiết kế kiến trúc phân tầng và mô hình lai; (3) Thu thập và thực nghiệm huấn luyện trên dữ liệu thực tế; (4) Đánh giá hiệu năng thông qua các chỉ số định lượng.

2.2.2. Thu thập và xử lý dữ liệu

Dữ liệu nghiên cứu được thu thập bằng Scraper API từ 239 trang Facebook công khai tại Việt Nam và đối soát với danh sách đen của dự án "Chống Lừa Đảo" (CLĐ) [14, 30]. Sau quá trình sàng lọc các mẫu nhiễu và chất lượng thấp (nội dung quá ngắn hoặc lỗi crawl), tập dữ liệu còn lại 201 mẫu chất lượng cao.

- *Gán nhãn:* Dữ liệu được gán nhãn đa lớp bao gồm: Lừa đảo việc làm (Job scam - 63 mẫu), Lừa đảo cờ bạc (Gambling scam - 72 mẫu) và An toàn (Safe/None - 66 mẫu).

- *Phân chia dữ liệu:* Để đảm bảo tính khách quan, nhóm tác giả sử dụng phương pháp chia tập phân tầng (Stratified Split), duy trì tỷ lệ phân bố các lớp tương đồng giữa các tập dữ liệu. Tập dữ liệu được chia theo tỷ lệ: huấn luyện (162 mẫu), kiểm định (29 mẫu) và kiểm thử độc lập (48 mẫu).

Nhận thấy số lượng 201 mẫu là tương đối nhỏ để huấn luyện tối ưu mô hình học sâu, nghiên cứu đã áp dụng các kỹ thuật tăng cường dữ liệu (Data Augmentation) cơ bản cho văn bản để làm phong phú tập huấn luyện. Đồng thời, sự chênh lệch nhỏ về số lượng giữa các lớp (do tính chất phân bố thực tế của dữ liệu) đã được xử lý thông qua việc điều chỉnh hàm mất mát có trọng số (Weighted Cross - Entropy Loss) trong quá trình huấn luyện, đảm bảo mô hình có tính tổng quát cao và không bị thiên lệch (bias) về lớp đa số.

2.2.3. Mô hình lai và quy trình trích xuất đặc trưng

Nghiên cứu đề xuất mô hình lai (Hybrid Model) kết hợp giữa học sâu và học máy truyền thống. Quy trình trích xuất đặc trưng được thực hiện song song:

- *Đặc trưng ngữ nghĩa*: Sử dụng PhoBERT-v2 để chuyển đổi nội dung văn bản thành vector nhúng có kích thước 768 chiều, giúp nắm bắt sắc thái ngôn từ thao túng tâm lý.

- *Đặc trưng hành vi*: Trích xuất các siêu dữ liệu (metadata) định lượng như lượt theo dõi, lượt tương tác, trạng thái xác thực.

Tại tầng kỹ thuật kết hợp đặc trưng (Feature Fusion), vector nhúng 768 chiều

từ PhoBERT được thực hiện phép nối ngang (concatenate/hstack) trực tiếp với vector lưu trữ các giá trị số của metadata hành vi. Vector tổng hợp duy nhất này sau đó mới được đưa vào thuật toán XGBoost để thực hiện phân loại đa lớp, tận dụng được khả năng xử lý xuất sắc của XGBoost đối với các điểm dữ liệu không đồng nhất.

2.2.4. Thiết kế kiến trúc hệ thống 5 tầng

Hệ thống được thiết kế theo kiến trúc phân tầng (layered architecture) nhằm đảm bảo tính linh hoạt và khả năng giải thích:

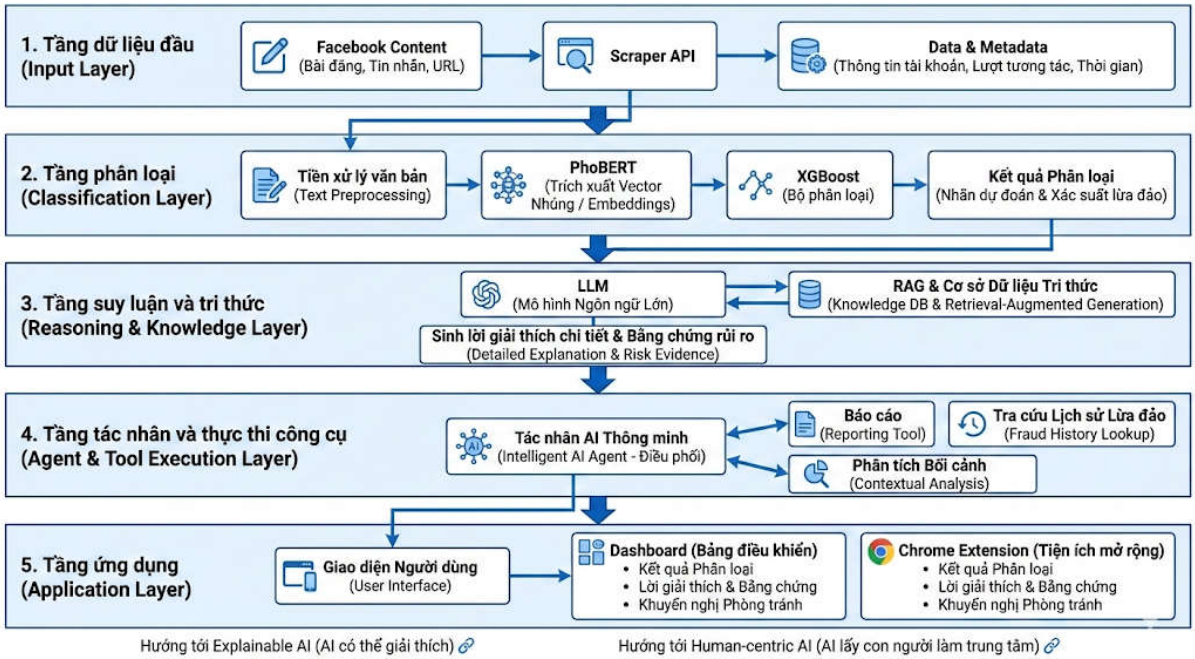
- *Tầng dữ liệu đầu vào:* Tiếp nhận URL, văn bản hoặc hình ảnh từ người dùng qua Scraper API.

- *Tầng phân loại (ML Layer):* Vận hành mô hình lai PhoBERT + XGBoost để đưa ra nhãn dự đoán tức thời.

- *Tầng suy luận và tri thức (Reasoning & Knowledge Layer):* Ứng dụng kỹ thuật RAG để truy xuất các văn bản Luật và kịch bản lừa đảo từ ChromaDB, giúp LLM sinh lời giải thích có căn cứ pháp lý.

- *Tầng tác nhân (Agent & Tool Layer):* Sử dụng giao thức MCP để điều phối các tác nhân AI gọi công cụ tra cứu danh sách đen số điện thoại và tài khoản ngân hàng theo thời gian thực.

- *Tầng ứng dụng:* Cung cấp giao diện tương tác qua Dashboard web và tiện ích Chrome Extension.



Hình 1: Sơ đồ kiến trúc hệ thống

2.2.5. Công cụ và môi trường thực nghiệm

Hệ thống được phát triển trên ngôn ngữ Python 3.11, sử dụng các framework hiện đại như FastAPI, LangChain và Next.js. Ngoài ra, hệ thống tích hợp giao diện từ Open WebUI nhằm hỗ trợ tương tác với mô hình ngôn ngữ lớn, hiển thị kết quả phân tích và cung cấp trải nghiệm sử dụng trực quan cho người dùng [19]. Toàn bộ giải pháp được đóng gói bằng Docker và Docker Compose, triển khai thực nghiệm trên máy chủ AWS EC2 (t3.large) để đo đạc độ trễ và hiệu năng thực tế. Hiệu quả của phương pháp được kiểm chứng qua các chỉ số Accuracy, Precision, Recall, F1-Score và diện tích dưới đường cong ROC (AUC).

3. Kết quả và thảo luận

3.1. Kết quả thực nghiệm mô hình phân loại

Hệ thống được thử nghiệm trên tập dữ liệu gồm 201 mẫu chất lượng cao sau khi đã loại bỏ các mẫu nhiễu. Để đánh giá khách quan, nhóm tác giả thực hiện thống kê phân bổ các lớp dữ liệu trước khi huấn luyện.

Bảng 1. Thống kê phân bổ nhãn trong tập dữ liệu

Loại nhãn	Giá trị nhãn	Ý nghĩa	Số lượng mẫu
is_scam	1	Trang lừa đảo	138
	0	Trang an toàn	101
scam_type	0	Lừa đảo việc làm (Job scam)	63
	1	Lừa đảo cờ bạc (Gambling scam)	72
	2	An toàn (Safe / None)	66
	-1	Dữ liệu chất lượng thấp	38

Quá trình huấn luyện mô hình lai PhoBERT-v2 + XGBoost cho thấy sự hội tụ ổn định. Thời gian tinh chỉnh (fine-tuning) PhoBERT-v2 mất khoảng 461 giây, trong khi giai đoạn huấn luyện bộ phân loại XGBoost chỉ mất 0,615 giây trên các đặc trưng đã trích xuất. Hiệu năng của mô hình đề xuất được so sánh với các phương pháp cơ sở (baseline) tại Bảng 2.

Bảng 2. Tổng hợp kết quả thực nghiệm các mô hình

Mô hình (Model)	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Inference Time
Các phương pháp cơ sở						
XGBoost (Metadata Only)	90,67 %	90 %	92,43 %	93,10 %	0,9402	0,018 ms
SVM + TF-IDF	87,50 %	82,35 %	100 %	90,32 %	0,9241	0,068 ms
Các mô hình Deep Learning đơn lẻ						
ViBERT	68,75 %	74,07 %	71,43 %	72,73 %	0,6741	281,61 ms
PhoBERT-base (Original)	66,67 %	65 %	92,86 %	76,47 %	0,7402	278,67 ms
PhoBERT-base-v2	75 %	73,53 %	89,29 %	80,65 %	0,7884	278,82 ms
Mô hình đề xuất						
PhoBERT-v2 + XGBoost	92,58 %	92,66 %	92,86 %	92,76 %	0,9509	0,022 ms

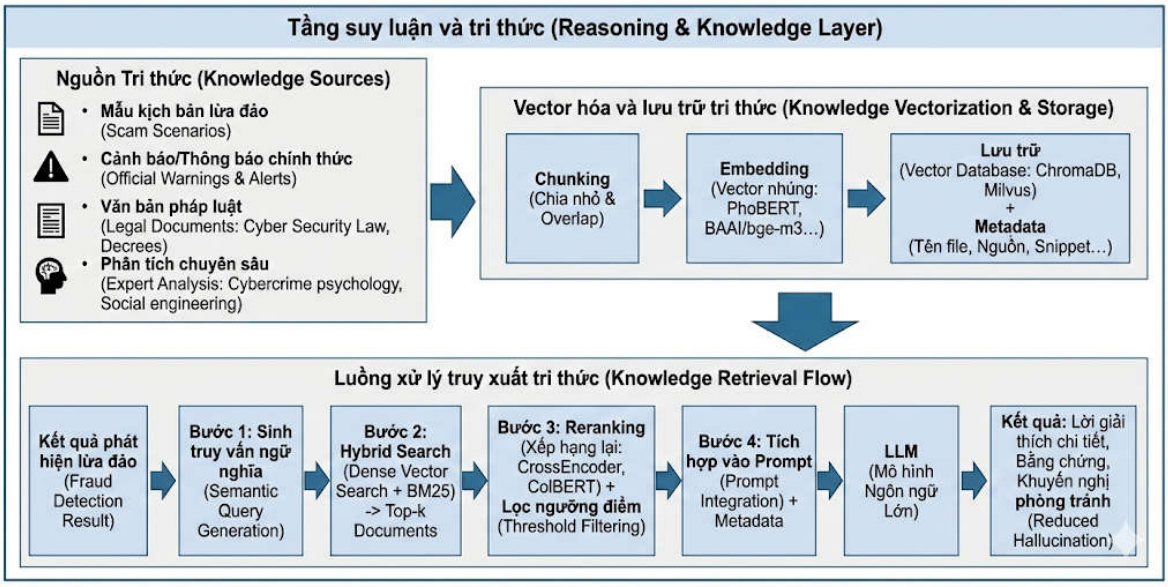
Nguồn: Nhóm tác giả tổng hợp từ thực nghiệm

Kết quả tại Bảng 2 minh chứng rằng việc kết hợp ngữ nghĩa chuyên sâu và đặc trưng hành vi giúp mô hình đề xuất đạt độ chính xác cao nhất và khả năng phân biệt (ROC - AUC) tốt nhất.

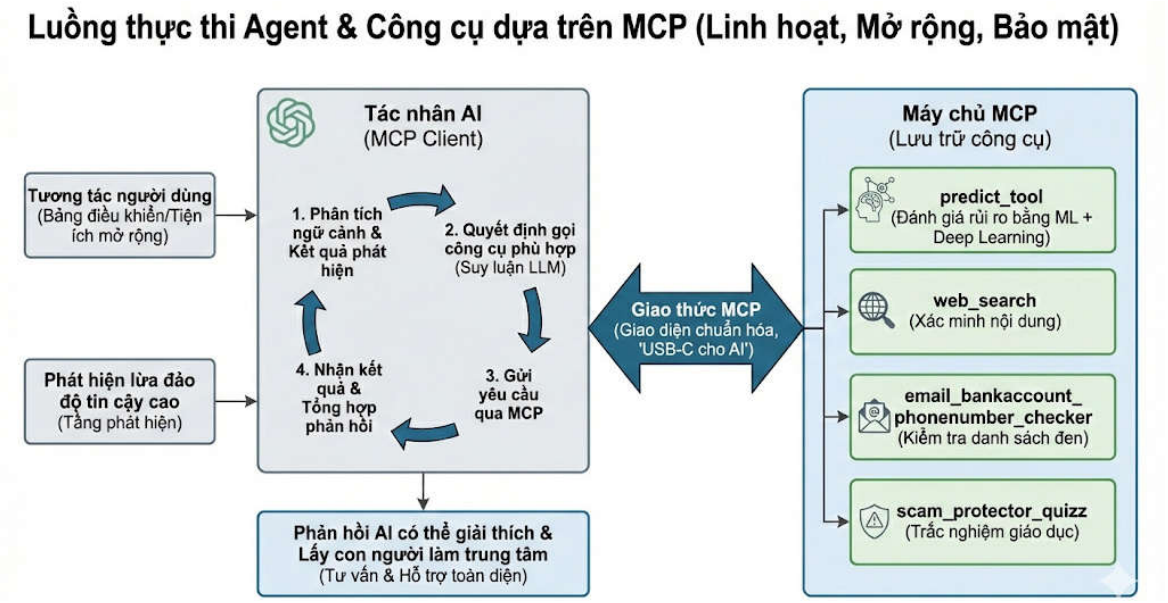
3.2. Hiệu quả của kỹ thuật RAG và giao thức MCP

Hệ thống thực nghiệm được cấu hình để kết nối với các mô hình ngôn ngữ lớn (LLM) tiên tiến thông qua giao diện trung gian Open WebUI (MCP Client) kết nối tới MCP Server. Trong cơ chế tăng cường truy xuất tri thức (RAG), tham số tìm kiếm ngữ nghĩa được thiết lập lấy top kết quả phù hợp nhất với K = 3. Thực nghiệm trên 50 câu hỏi kiểm thử cho thấy kỹ thuật RAG giúp đạt độ chính xác ngữ cảnh là 0,88. Chỉ số này được đo lường chặt chẽ dựa trên tỷ lệ các đoạn văn bản được hệ thống truy xuất từ cơ sở dữ liệu Vector (ChromaDB) thực sự liên quan

đến nội dung câu hỏi và hỗ trợ LLM trả lời đúng, không bị lạc đề. Đồng thời, AI Agent tích hợp qua giao thức MCP đạt tỷ lệ gọi công cụ tra cứu thành công 96,25 % với thời gian phản hồi trung bình khoảng 3,8 giây cho các tác vụ kiểm tra số điện thoại và tài khoản ngân hàng lừa đảo theo thời gian thực.



Hình 2: Sơ đồ luồng xử lý dữ liệu RAG



Hình 3: Sơ đồ luồng thực thi của AI Agent và công cụ dựa trên MCP

3.3. Thảo luận

3.3.1. So sánh với các nghiên cứu trong nước

Dự án "Chống Lừa Đảo" (CLĐ) của Ngô Minh Hiếu (2020): Cần đối chiếu để thấy rằng trong khi CLĐ chủ yếu dựa trên tập luật (rule-based) và báo cáo từ cộng đồng (crowdsourcing) để chặn URL, hệ thống của nhóm đã tiến xa hơn bằng cách sử dụng LLM để giải thích lý do lừa đảo và tư vấn tâm lý thao túng [30].

Nghiên cứu về PhoBERT của Nguyễn Quốc Đạt và cộng sự (2020): So sánh hiệu năng xử lý tiếng Việt. Nhóm có

thể biện luận rằng nghiên cứu này không chỉ dừng lại ở mức phân loại văn bản đơn thuần như các ứng dụng PhoBERT trước đó đã tích hợp thành một hệ thống đa lớp (Hybrid Model) kết hợp với đặc trưng hành vi (metadata) để đạt độ chính xác cao hơn (92,58 %) [4].

Các nghiên cứu về Retrieval - Augmented Generation của Amugongo và cộng sự (2025) và FPT Cloud (2025) cho thấy việc bổ sung tầng truy xuất tri thức giúp tăng cường khả năng suy luận của mô hình ngôn ngữ lớn [11, 12]. Bên cạnh đó, các nghiên cứu về AI Agents của Wang và cộng sự (2023), Fang và cộng sự (2025), Sapkota và cộng sự (2025) nhấn mạnh vai trò của tầng tác nhân trong việc tự động hóa và xử lý dữ liệu phức tạp [7, 8, 9]. Kế thừa các hướng tiếp cận này, hệ thống đề xuất được mở rộng với tầng suy luận (RAG) và tầng tác nhân nhằm phù hợp với bài toán dữ liệu mạng xã hội.

3.3.2. So sánh với các nghiên cứu ngoài nước

Nghiên cứu của Ronish Nair và cộng sự (2025) về mô hình lai (PhishEmailLLM) kết hợp LLM và XGBoost cho email [33]. Cần so sánh để khẳng định giải pháp của mình đã tối ưu hóa cho ngôn ngữ tiếng Việt và môi trường Facebook, đồng thời bổ sung kỹ thuật RAG để giảm "ảo giác" mà nghiên cứu của Nair chưa tập trung sâu.

Nghiên cứu của Matthew Benjamin (2025), đối chiếu về khả năng xử lý thời gian thực. Trong khi Benjamin tập trung vào các thách thức về độ trễ (latency), hệ thống đã chứng minh được thời gian phản hồi trung bình dưới 4 giây dù tích hợp các quy trình phức tạp như RAG và MCP [29].

Nghiên cứu của René Meléndez và cộng sự (2024), so sánh về ưu thế của mô hình Transformer so với các mô hình truyền thống (SVM, Naive Bayes) [32]. Kết quả của nhóm nghiên cứu (Bảng 2) củng cố thêm nhận định của Meléndez nhưng cho thấy sự kết hợp Transformer + XGBoost mang lại hiệu quả vượt trội hơn so với việc chỉ dùng Transformer đơn lẻ.

3.3.3. Phân tích và biện luận kết quả

Dựa trên các bảng số liệu thực nghiệm, nghiên cứu rút ra các biện luận khoa học sau:

- *Tầm quan trọng của Metadata:* Siêu dữ liệu hành vi là tín hiệu cực kỳ mạnh mẽ. Mô hình chỉ dùng Metadata đã đạt Accuracy 90,67 %, trong khi các mô hình chỉ dựa trên văn bản (PhoBERT thuần túy) đạt dưới 75 % do nội dung lừa đảo ngày càng tinh vi [34].

- *Sự cân bằng giữa Precision và Recall:* Mô hình SVM đạt Recall 100 % nhưng Precision thấp (82,35 %), dẫn đến nhiều báo động giả. Mô hình đề xuất đạt Precision 92,66 %, giúp giảm thiểu phiền hà cho người dùng.

- *Giải quyết bài toán "hộp đen" của AI:* Việc tích hợp RAG và MCP buộc hệ thống phải giải thích dựa trên bằng chứng pháp lý (như Luật An ninh mạng) và dữ liệu thời gian thực (danh sách đen số điện thoại/số tài khoản), biến công cụ phân loại thành một trợ lý tư vấn minh bạch [28, 31].

4. Kết luận và đề xuất

4.1. Kết luận

Nghiên cứu đã xây dựng thành công hệ thống nhận diện lừa đảo Facebook với kiến trúc phân tầng linh hoạt, đạt độ chính xác 92,58 %. Hệ thống không chỉ là một

công cụ lọc chặn mà còn là một trợ lý ảo hỗ trợ nâng cao nhận thức an toàn thông tin cho người dùng thông qua các tính năng tư vấn và trắc nghiệm tương tác.

Tính minh bạch (Explainable AI): Khẳng định sự khác biệt so với các mô hình "hộp đen" chỉ đưa ra nhãn cảnh báo mà không có giải thích pháp lý như hệ thống của nhóm (độ chính xác ngữ cảnh RAG đạt 0,88).

Khả năng tương tác (MCP Agent): Đây là điểm mới so với hầu hết các nghiên cứu hiện tại vốn chưa tích hợp giao thức MCP để tra cứu danh sách đen thời gian thực một cách tự động.

4.2. Đề xuất và hướng phát triển

Trong tương lai, nhóm nghiên cứu đề xuất mở rộng khả năng xử lý đa phương thức (Multimodal AI) để nhận diện hình ảnh Deepfake và hóa đơn giả mạo. Đồng thời, nghiên cứu ứng dụng các kỹ thuật lượng tử hóa để triển khai mô hình trực tiếp trên thiết bị người dùng (On-device AI) nhằm đảm bảo quyền riêng tư.

Lời cảm ơn: Nghiên cứu này được hỗ trợ và sử dụng số liệu từ đề tài nghiên cứu khoa học sinh viên Trường Đại học Tài nguyên và Môi trường Hà Nội năm học 2025 - 2026. Mã số: HUNRE. NH.2026.11.21.

TÀI LIỆU THAM KHẢO

[1]. Sơn Vân (2025). *Facebook thành điểm nóng lừa đảo: Việt Nam và Singapore yêu cầu Meta hành động*. Hà Nội.

[2]. Quỳnh Dương (2025). *Những kẻ lừa đảo 'thao túng tâm lý' nạn nhân thế nào*?. VNExpress.

[3]. Cole Stryker, Jim Holdsworth (2024). *What is NLP (Natural Language Processing)?*. IBM.

[4]. Dat Quoc Nguyen, Anh Tuan Nguyen (2020). *PhoBERT: Pre-trained language models for Vietnamese*. Cornell University. https://doi.org/10.48550/arXiv.2003.00744.

[5]. Candice Bentéjac, Anna Csörgő, Gonzalo Martínez-Muñoz (2019). *A Comparative Analysis of XGBoost.* Cornell University. https://doi.org/10.48550/arXiv.1911.01914.

[6]. Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, Ajmal Mian (2024). *A Comprehensive Overview of Large Language Models*. Cornell University. https://doi.org/10.48550/arXiv.2307.06435.

[7]. Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, Ji-Rong Wen (2024). *A survey on large language model based autonomous agents*. Frontiers of Computer Science. https://doi.org/10.1007/s11704-024-40231-1.

[8]. Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, Zaiqiao Meng (2025). *A Comprehensive Survey of Self-Evolving AI Agents: A New Paradigm Bridging Foundation Models and Lifelong Agentic Systems*. Cornell University https://doi.org/10.48550/arXiv.2508.07407.

[9]. Ranjan Sapkota, Konstantinos I. Roumeliotis, Manoj Karkee (2026). *AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges*. Information Fusion 126 (2026) 103599.

[10]. Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, Diyi Yang (2025). *Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce.* Cornell University. https://doi.org/10.48550/arXiv.2506.06576.

[11]. Lameck Mbangula

Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, Jan Seidel (2025). *Retrieval augmented generation for large language models in healthcare: A systematic review.* PLOT Digital Health. Doi: 10.1371/journal.pdig.0000877.

[12]. FPT Cloud (2025). *Tối ưu hóa sức mạnh của AI tạo sinh cùng Retrieval-Augmented Generation (RAG).* https://fptcloud.com/toi-uu-hoa-suc-manh-cua-ai-tao-sinh-cung-retrieval-augmented-generation-rag/.

[13]. Anna Gutowska (2024). *What is Model Context Protocol (MCP)?*. IBM.

[14]. Krasnoludkolo (2026). *Facebook Scraper*. RapidAPI. Available: https://rapidapi.com/krasnoludkolo/api/facebook-scraper3.

[15]. Felipe Farias, Teresa Ludermir, Carmelo Bastos-Filho (2020). *Similarity Based Stratified Splitting: an approach to train better classifiers*. Cornell University. https://doi.org/10.48550/arXiv.2010.06099.

[16]. Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, Rajiv Ranjan (2023). *Explainable AI (XAI): Core Ideas, Techniques, and Solutions*. ACM Computing Surveys, vol. 55, no. 9, p. 1 - 33. https://doi.org/10.1145/3561048.

[17]. Davor Horvatić, Tomislav Lipic (2021). *Human-Centric AI: The Symbiosis of Human and Artificial Intelligence*. Entropy, vol. 23, no. 3, p. 23.

[18]. Tomas Bueno Momcilovic, Dian Balta, Beat Buesser, Giulio Zizzo, Mark Purcell (2024). *Knowledge-Augmented Reasoning for EUAIA Compliance and Adversarial Robustness of LLMs*. Cornell University. https://doi.org/10.48550/arXiv.2410.09078.

[19]. Open WebUI (2026). *Open WebUI: Self-Hosted AI Platform*. Open WebUI. https://openwebui.com/.

[20]. Trọng Đạt (2025). *Facebook gỡ 5,4 triệu bài đăng lừa đảo, gian lận tại Việt Nam*. VNEXPRESS.

[21]. Stuart J. Russell, Peter Norvig (2002). *Artificial Intelligence: A modern*. Approach, Prentice-Hall.

[22]. Đinh Mạnh Tường (2006). *Giáo trình trí tuệ nhân tạo*. Đại học Quốc gia Hà Nội.

[23]. Bộ Khoa học và Công nghệ (2023). *Cảnh báo 16 hình thức lừa đảo phổ biến trên không gian mạng*.

[24]. Đinh Minh Hòa, Phạm Ngọc Bảo, Huỳnh Vũ Lê, Lê Huỳnh Nghiêm, Nguyễn Thị Thúy A, Trần Khải Thiện (2024). *Mô hình ngôn ngữ lớn và ứng dụng*. HUFLIT Journal of Science. Trường Đại học Ngoại ngữ - Tin học Thành phố Hồ Chí Minh.

[25]. Federal Trade Commission (2022). *How To Recognize and Avoid Phishing Scams*. https://consumer.ftc.gov/articles/how-recognize-avoid-phishing-scams.

[26]. Gemma Martin (2024). *Machine learning and AI in fraud detection: How businesses should be harnessing Machine Learning and AI for advanced fraud detection*. Experian.

[27]. J. Wilk-Jakubowski, L. Pawlik, Grzegorz Wilk-Jakubowski, Aleksandra Sikora (2025). *Machine Learning and Neural Networks for Phishing Detection: A Systematic Review (2017 - 2024)*. Electronics, vol. 14, issue 18. https://doi.org/10.3390/electronics14183744.

[28]. Liming Jiang (2024). *Detecting Scams Using Large Language Models*. Cornell University. https://doi.org/10.48550/arXiv.2402.03147.

[29]. Matthew Benjamin (2025). *Real-Time Fraud Detection Systems: Challenges and Implementation Using Machine Learning*. https://www.researchgate.net/publication/392263683_Real-Time_Fraud_Detection_Systems_Challenges_and_Implementation_Using_Machine_Learning.

[30]. Ngô Minh Hiếu (2020). *Chống lừa đảo*. Công ty TNHH Doanh nghiệp xã hội chống lừa đảo. https://chongluadao.vn/posts/about-us.

[31]. Quy Quach Phu, Tuyen Dang Trong, Hoang Phan Huy, Trung Nguyen Quoc, Vinh Truong Hoang, Meirambek Zhaparov (2024). *An Improment of Large Language Model for Vietnamese Traffic Law Question Answer System*. International Conference on Intelligence of Things, Hanoi. FPT University.

[32]. René Meléndez, Michal Ptaszynski, Masui Fumito (2024). *Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection*. Electronics, 13 (24), 4877. https://doi.org/10.20944/preprints202410.1467.v1.

[33]. Ronish Nair, Fahim Abbasi, Shahbaz Pervez (2025). *PhishEmailLLM: A Meta Model Approach to Detect Phishing emails by leveraging LLMs and Machine Learning models*. ACSW '25: Proceedings of the 2025 Australasian Computer Science Week. https://dl.acm.org/doi/proceedings/10.1145/3727166.

[34]. Thuan Nguyen Hieu, Hieu Cao Nguyen Minh, Hung To Van, Bang Vo Quoc (2020) *ReINTEL Challenge 2020: Vietnamese Fake News Detection using Ensemble Model with PhoBERT embeddings.* In Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, pages 1 - 5, Hanoi, Vietnam. Association for Computational Linguistics. https://aclanthology.org/2020.vlsp-1.1.pdf.