Đánh giá độ tin cậy của phần mềm phát hiện AI

Evaluate the reliability of AI detection software

Trần Quang Cảnh[1,2*], Hoàng Thị Chỉnh[1,2]

[1]Trường Đại học Kinh tế - Luật, Thành phố Hồ Chí Minh, Việt Nam
[2]Đại học Quốc gia Thành phố Hồ Chí Minh, Thành phố Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ, Email: tranquangcanh3110@gmail.com

THÔNG TIN

DOI:10.46223/HCMCOUJS.tech.vi.20.1.3589.2025

Ngày nhận: 26/07/2024

Ngày nhận lại: 01/10/2024

Duyệt đăng: 11/10/2024

Từ khóa:

hiệu quả phần mềm phát hiện văn bản do AI tạo sinh; phân tích ROC; văn bản AI tạo sinh

TÓM TẮT

Nghiên cứu này đã đánh giá hiệu quả của phần mềm phát hiện AI trong việc phân biệt giữa văn bản do con người tạo ra và AI, sử dụng mô hình Vistral-7B-Chat. Đánh giá bao gồm 30 công cụ phát hiện được thử nghiệm trên 10 mẫu văn bản, được phân chia đều giữa nguồn người và AI. Thống kê mô tả và phân tích đường cong ROC đã được sử dụng để đánh giá độ chính xác của các công cụ này. Các phát hiện cho thấy phần mềm phân biệt hiệu quả giữa AI và văn bản do con người tạo ra, đạt được điểm AUC là 1, biểu thị độ chính xác gần như hoàn hảo. Nghiên cứu đã xác định sự thay đổi trong hiệu suất của công cụ, nhấn mạnh sự cần thiết của các cải tiến liên tục để giải quyết các thách thức diễn giải và lảng tránh. Nghiên cứu này góp phần vào sự hiểu biết về phát hiện văn bản AI, nhấn mạnh nhu cầu cấp bách về các công cụ mạnh mẽ để bảo vệ tính toàn vẹn của nội dung do con người tạo ra khi các công nghệ AI tiến bộ.

Keywords:

effective software detection of documents by AI born; analyzing ROC; AI text created

ABSTRACT

This study examined how well AI detection software can distinguish between human-generated and AI-generated text using the Vistral-7B-Chat model. The evaluation included 30 detection tools evaluated on 10 text samples, split evenly between human and AI sources. Using descriptive statistics and ROC curve analysis, the study measured the accuracy of these tools. The results showed that the software successfully distinguished between AI and human-generated text, achieving an AUC score of 1, indicating near-perfect accuracy. The study noted variability in performance across tools, highlighting the need for continuous improvement to address interpretation and evasion challenges. This research advances our understanding of AI text detection, highlighting the urgency of robust tools to protect the integrity of human-generated content as AI technologies evolve.

1. Giới thiệu

Trong xã hội đương đại, tầm quan trọng của việc tạo nội dung được nâng cao do những tiến bộ trong trí tuệ nhân tạo (AI) và tạo văn bản tự động. Sự xuất hiện của các mô hình ngôn ngữ tinh vi như GPT-4 đã làm dấy lên những lo ngại về độ chính xác và độ tin cậy của thông tin.

Để đối phó với những thách thức này, các công cụ đã được phát triển để phân biệt giữa nội dung do con người tạo ra và nội dung do AI tạo ra. Tuy nhiên, hiệu quả của các phương pháp phát hiện này đòi hỏi phải đánh giá toàn diện trong các môi trường thực tế.

Cuộc điều tra này sử dụng phân tích đường cong ROC để đánh giá độ tin cậy của các hệ thống nhận dạng văn bản AI. Đường cong ROC đóng vai trò như một công cụ thiết yếu để đánh giá hiệu suất của các mô hình phân loại, cung cấp thông tin chi tiết về độ chính xác của các cơ chế phát hiện trong phần mềm. Nghiên cứu tập trung vào hai loại nội dung chính: văn bản do con người viết (HM) và văn bản do AI tạo ra (Vistral-7B-Chat - AI). Bằng cách đối chiếu các hình thức nội dung này, nghiên cứu đã tìm cách đánh giá khả năng của phần mềm phát hiện AI để phân biệt giữa các nguồn gốc văn bản khác nhau.

Việc kiểm tra ba mươi công cụ để phát hiện văn bản do AI tạo ra, được đánh giá trên hai loại nội dung riêng biệt, làm sáng tỏ khả năng và hạn chế của phần mềm phát hiện AI đương đại. Nghiên cứu này đánh giá độ tin cậy của các ứng dụng bằng cách định lượng mức độ thành thạo của chúng trong việc xác định chính xác tài liệu. Các phát hiện tiết lộ rằng các công cụ này thể hiện hiệu suất trung bình đáng khen ngợi trong việc phân biệt giữa văn bản do AI tạo ra và văn bản do con người tạo ra. Mặc dù sự khác biệt quan sát được là nhỏ và thiếu ý nghĩa thống kê trong phân tích ROC, kết quả ngụ ý rằng phần mềm xác định nội dung do con người tạo ra hiệu quả hơn so với văn bản do AI tạo ra. Điều này cho thấy khả năng phát hiện của các công cụ tương đối cân bằng cho cả hai loại văn bản, phù hợp với mục tiêu của các phương pháp phát hiện văn bản AI đương đại. Một phát hiện đáng chú ý là hiệu suất nâng cao của các phương pháp kết hợp, chỉ ra rằng việc tích hợp nhiều kỹ thuật nhận dạng và tổng hợp dữ liệu có thể cải thiện rõ rệt độ chính xác.

Những phát hiện này nâng cao sự hiểu biết về phần mềm phát hiện AI và thiết lập cơ sở cho các cải tiến trong tương lai. Ý nghĩa mở rộng đến các ứng dụng thực tế, nhấn mạnh tầm quan trọng của tính toàn vẹn và tính xác thực của dữ liệu. Những tiến bộ liên tục trong công nghệ phát hiện AI sẽ thúc đẩy việc sử dụng AI có trách nhiệm, bảo vệ người tiêu dùng khỏi thông tin sai lệch.

2. Các nghiên cứu trước

Sự xuất hiện của các mô hình ngôn ngữ lớn đã làm nổi bật thách thức trong việc phát hiện nội dung do AI tạo ra (Wang, 2024). Các cơ chế phát hiện AI hiện tại thể hiện hiệu quả hạn chế (Bellini & ctg., 2024; Legaspi & ctg., 2024), đặc biệt chống lại các chiến lược trốn tránh làm ảnh hưởng đến độ chính xác (Kumarage & ctg., 2023; Krishna & ctg., 2023; Lu & ctg., 2023; Sadasivan & ctg., 2023). Các mô hình ngôn ngữ có thể được thao tác thông qua các lời nhắc cụ thể để bỏ qua các hệ thống phát hiện (Kumarage & ctg., 2023; Lu & ctg., 2023). Mặc dù gặp khó khăn với các mô hình tiên tiến như GPT-4 (Walters, 2023), các công cụ như Copyleaks, Turnitin và Originality.ai phân biệt hiệu quả văn bản do AI tạo ra với chữ viết của con người. Hiệu suất của các công cụ phát hiện AI phụ thuộc vào các đặc điểm văn bản cụ thể, với độ chính xác dao động trong khoảng 55.29% đến 97.0% (Akram, 2023). Đáng chú ý, các kỹ thuật học tập đối nghịch sử dụng RADAR hoạt động vượt trội trong việc nhận biết nội dung do AI tạo ra giữa những thách thức diễn giải (Hu & ctg., 2023). Phương pháp tìm kiếm ngữ nghĩa nâng cao độ chính xác của phát hiện, ngay cả trong các kịch bản diễn giải (Krishna & ctg., 2023). Nghiên cứu thực nghiệm, được củng cố bởi các bộ dữ liệu mở rộng và các phương pháp phát hiện khác nhau, cho thấy rằng phát hiện văn bản AI có thể đạt đến mức độ chính xác đáng kể (Chakraborty & ctg., 2023). Các giải pháp như Ghostbuster giải quyết các kỹ thuật trốn tránh và khẳng định độ chính xác vượt trội trên các lĩnh vực khác nhau (Verma & ctg., 2023).

Bất chấp sự gia tăng của các hệ thống phát hiện AI, các nghiên cứu thực nghiệm cho thấy độ tin cậy của chúng vẫn bị hạn chế, đặc biệt là liên quan đến các diễn giải có nhiều sắc thái. Trong khi các chiến lược nâng cao như tìm kiếm ngữ nghĩa và học đối nghịch cho thấy lời hứa về độ chính xác nâng cao, nghiên cứu sâu hơn là bắt buộc để thiết lập các công cụ đáng tin cậy hơn phù hợp với các bối cảnh đa dạng. Nghiên cứu này cung cấp dữ liệu thực nghiệm để đánh giá hiệu quả của các hệ thống nhận dạng văn bản do AI tạo ra trong bối cảnh ngôn ngữ tiếng Việt.

3. Phương pháp nghiên cứu

3.1. Thiết kế nghiên cứu

Nghiên cứu này đã đánh giá 30 công cụ phần mềm phát hiện AI (chi tiết trong phụ lục (bản online)). Đánh giá đã sử dụng các mẫu văn bản được xác định trước, bao gồm nội dung do con người viết và Vistral-7B-Chat. Phương pháp này tạo điều kiện cho việc đánh giá khách quan về sự thành thạo của các chương trình phát hiện AI trong việc phân biệt giữa văn bản do con người tạo ra và AI. Tập dữ liệu bao gồm mười mẫu văn bản - năm mẫu do AI tạo ra và năm mẫu do con người tạo ra. Văn học nhân văn bao gồm các chủ đề đa dạng, bao gồm hai đoạn từ sách giáo khoa kinh tế và quản lý tiền AI và ba bản tóm tắt từ các tạp chí học thuật. Vistral-7B-Chat đã tạo ra nội dung AI bằng cách sao chép các đoạn này. Thiết kế nghiên cứu này đảm bảo tính nhất quán giữa AI và văn bản của con người, do đó nâng cao độ tin cậy của kết quả phân tích. Các mẫu văn bản được sử dụng trong nghiên cứu này được nêu dưới đây. Đoạn văn được viết bởi một con người:

Đoạn 1: “Nho giáo ủng hộ một hệ thống “đại diện,” trong đó các bà vợ được mong đợi sẽ tích cực tham gia, trong khi các ông chồng giữ quyền lực tối thượng. Tuy nhiên, khái niệm “vợ phụ thuộc” của Nho giáo mâu thuẫn với quan điểm của người Việt Nam về hôn nhân. Mặc dù Nho giáo có ảnh hưởng đáng kể đến tư tưởng trí thức Việt Nam, đặc biệt là trong thời kỳ phong kiến, nó cũng duy trì những tư tưởng tiến bộ về vai trò của phụ nữ trong gia đình. Ở Việt Nam, hôn nhân được xem là quan hệ đối tác với trách nhiệm chung, như được minh họa bằng câu nói “phụ nữ quản lý gia đình, đàn ông xử lý công việc đối ngoại.” Nghiên cứu này xem xét cách Nho giáo thể hiện trong các cuộc hôn nhân Việt Nam đương đại. Các nhà nghiên cứu đã sử dụng phân tích thống kê mô tả, phân tích yếu tố thăm dò (EFA) và mô hình hỗn hợp Gaussian (GMM) để phát triển một hệ phương trình đồng thời. Hệ thống này đánh giá tác động của Nho giáo đối với cấu trúc hôn nhân Việt Nam bằng cách kiểm tra động lực quyền lực vợ chồng trong cả lĩnh vực kinh tế và ra quyết định. Mặc dù ít nổi bật hơn so với thời kỳ phong kiến và đầu những năm 1900, những phát hiện chỉ ra rằng Nho giáo tiếp tục định hình các mối quan hệ gia đình ngày nay. Trong khi các giá trị Nho giáo vẫn còn phù hợp trong một số hộ gia đình Việt Nam, sự tiến bộ của xã hội công nghiệp hiện đại đã thay đổi và tác động đến các chuẩn mực xã hội”.

Đoạn 2: “Trong ba thập kỷ qua, Việt Nam đã trải qua một sự chuyển đổi kinh tế đáng kể, vươn lên từ một trong những nước nghèo nhất thế giới để đạt được vị thế thu nhập trung bình thấp. Bất chấp những thách thức do đại dịch Covid-19 và thiên tai gây ra, quốc gia này vẫn duy trì tốc độ tăng trưởng kinh tế trung bình mạnh mẽ từ năm 2016 đến năm 2020, vượt trội so với nhiều đối tác trong khu vực và toàn cầu. Trong khi khu vực dịch vụ đã trải qua sự mở rộng đáng kể, các lĩnh vực khác, đặc biệt là nông nghiệp, lâm nghiệp và thủy sản, đã chứng kiến sự suy giảm. Nghiên cứu này sử dụng mô hình ARIMA để dự báo cơ cấu kinh tế Việt Nam đến năm 2025, với dữ liệu từ năm 1987 đến năm 2019, không bao gồm thuế sản phẩm nhưng bao gồm trợ cấp sản phẩm. Phương pháp này bao gồm đánh giá tính ổn định chuỗi thời gian, ước tính các chức năng tự tương quan, xác định các tham số mô hình ARIMA, xác nhận các giả định và dự báo số liệu kinh tế cho năm 2025. Các dự báo của nghiên cứu cho thấy sự thay đổi đáng kể trong

bối cảnh kinh tế từ năm 2020 đến năm 2025, với sự chuyển dịch từ nông, lâm nghiệp và thủy sản sang dịch vụ. Đóng góp của công nghiệp và xây dựng dự kiến sẽ giảm từ 38% xuống 36%, trong khi khu vực dịch vụ được dự đoán sẽ tăng từ 46% lên 52%. Tỷ trọng nông, lâm nghiệp và thủy sản được dự đoán sẽ giảm từ 16% năm 2019 xuống còn khoảng 12% vào năm 2025. Những phát hiện này cung cấp những hiểu biết định lượng có giá trị ở cả cấp độ kinh tế vĩ mô và kinh tế vi mô, cần thiết để xây dựng và thực hiện các chính sách hiệu quả" (bài báo).

Đoạn 3: "Nghiên cứu này điều tra các số liệu tài chính ảnh hưởng như thế nào đến hiệu suất quản lý, được đo lường bằng Lợi nhuận trên Tài sản (ROA), giữa các công ty trên thị trường chứng khoán Việt Nam. Nghiên cứu phân tích dữ liệu từ 621 công ty trong lĩnh vực hàng tiêu dùng và công nghiệp năm 2019, sử dụng các phương trình đồng thời và phương pháp hồi quy Khoảnh khắc Tổng quát (GMM). Các phát hiện cho thấy ba yếu tố tài chính ảnh hưởng đáng kể đến hiệu suất quản lý: Vòng quay tài sản thể hiện ảnh hưởng mạnh nhất, tiếp theo là Tỷ suất lợi nhuận ròng và Tỷ lệ nợ trên tài sản. Hơn nữa, tỷ lệ nợ trên tài sản được tìm thấy có tác động đến tỷ suất lợi nhuận ròng. Những kết quả này khác biệt so với nghiên cứu trước đây và quan điểm của Hiệp hội Giao dịch Chứng khoán Việt Nam, cho thấy ảnh hưởng tối thiểu của vòng quay tổng tài sản đến hiệu quả quản lý. Rút ra từ những phát hiện này, các nhà nghiên cứu đưa ra các gợi ý để nâng cao hiệu quả quản trị doanh nghiệp, hỗ trợ các nhà đầu tư đánh giá hiệu quả quản lý khi xây dựng chiến lược đầu tư" (bài báo).

Đoạn 4: "Các nghiên cứu ban đầu ở Hawthorne, Hoa Kỳ, cho thấy mức độ ánh sáng tăng lên có thể nâng cao hiệu quả của công nhân. Các nhà nghiên cứu đã chọn hai nhóm công nhân: một trong điều kiện ánh sáng bình thường và một nhóm khác tiếp xúc với cường độ ánh sáng khác nhau. Nhóm với ánh sáng thay đổi cho thấy năng suất được cải thiện, hỗ trợ giả thuyết ban đầu. Đáng ngạc nhiên, nhóm có ánh sáng không thay đổi cũng chứng minh năng suất tăng lên. Kết quả bất ngờ này đã thúc đẩy điều tra thêm. Các nhà nghiên cứu kết luận rằng, ngoài những tiến bộ công nghệ, hành vi của con người đóng một vai trò quan trọng, dẫn đến sự tham gia của Mayo và nhóm của ông. Mayo đã tiến hành các thí nghiệm kéo dài mười tám tháng, tập trung vào phụ nữ lắp ráp rơle điện thoại. Nhóm của ông đã thực hiện nhiều cải tiến khác nhau đối với điều kiện làm việc, bao gồm chính sách nghỉ phép mới, kiểm tra nhà máy và giảm giờ làm việc. Những thay đổi này dẫn đến tăng năng suất đáng kể. Để xác nhận những phát hiện của họ, các nhà nghiên cứu đã quay trở lại điều kiện làm việc ban đầu, hy vọng năng suất sẽ giảm. Tuy nhiên, trái với dự đoán của họ, sản lượng tiếp tục tăng đều đặn" (giáo trình).

Đoạn 5: "Các giai đoạn phân tích ban đầu đòi hỏi phải kiểm tra kỹ lưỡng và tích hợp các kết quả tính toán tương ứng với các mục tiêu cụ thể. Cách tiếp cận này làm nổi bật những thách thức chính, đảm bảo sự tập trung của tổ chức và thiết lập một khuôn khổ cơ bản cho nghiên cứu trong tương lai. Sự thay đổi dữ liệu có thể là kết quả của tác động trực tiếp của các yếu tố đóng góp; vì vậy, điều cần thiết là phải xác định mức độ của các yếu tố này và nguyên nhân cơ bản của những thay đổi. Phân tích kinh doanh không chỉ đánh giá hiệu quả hoạt động tổng thể mà còn cho thấy các cơ hội tiềm năng để khai thác và xác định điểm yếu, từ đó hướng dẫn xây dựng các chiến lược tăng lợi nhuận và giảm tác động tiêu cực. Phương pháp này cho phép các tổ chức xác định chính xác các lĩnh vực cần cải thiện trong việc thực hiện và lý do chênh lệch, giúp họ nhận ra các cơ hội thích ứng tiềm năng. Khi được thực hiện hiệu quả, đánh giá này cho phép điều chỉnh chiến lược tổ chức và lập kế hoạch cho các mục tiêu trong tương lai" (giáo trình).

Văn bản do Vistral-7B-Chat viết: Đoạn 1: "Quá trình phân tích ban đầu đòi hỏi phải kiểm tra toàn diện và tích hợp các kết quả tính toán liên quan đến các mục tiêu ước tính cụ thể. Phương pháp này giải quyết các vấn đề chính, đòi hỏi sự tập trung của công ty và thiết lập nền tảng cho các giai đoạn nghiên cứu tiếp theo. Các biến thể dữ liệu có thể được quy cho các tác

động tức thời của nhiều yếu tố; do đó, điều cần thiết là phải xác định mức độ của những ảnh hưởng này và nguyên nhân cơ bản của sự dao động. Phân tích kinh doanh không chỉ đánh giá hiệu suất tổng thể mà còn cho thấy những lợi thế tiềm năng để tận dụng và xác định điểm yếu, từ đó thông báo cho việc phát triển các chiến lược để nâng cao điểm mạnh và giảm thiểu tác động tiêu cực. Cách tiếp cận này cho phép các tổ chức nhận ra những cải tiến có thể có trong việc thực hiện và hiểu lý do đằng sau sự khác biệt, do đó hướng dẫn họ dự tính các sửa đổi cần thiết. Khi được thực hiện thành công, đánh giá này tạo điều kiện cho việc tăng cường chiến lược tổ chức và xây dựng kế hoạch cho những nỗ lực trong tương lai".

Đoạn 2: "Nền kinh tế Việt Nam đã trải qua sự mở rộng phi thường trong 30 năm qua, nhanh chóng phát triển thành một quốc gia có thu nhập thấp. Từ năm 2016 đến năm 2020, Việt Nam đã đạt tốc độ tăng trưởng cao nhất toàn cầu và trong khu vực, ngay cả khi đối mặt với những trở ngại như đại dịch Covid-19 và thiên tai vào năm 2020. Cơ cấu kinh tế đã trải qua một sự chuyển đổi, với sự giảm tỷ trọng nông, lâm nghiệp và thủy sản, trong khi các lĩnh vực khác đã chứng kiến sự đóng góp tăng lên. Nghiên cứu này sử dụng dữ liệu từ nông, lâm, thủy sản, công nghiệp, xây dựng và dịch vụ từ năm 1987 đến năm 2019, không bao gồm thuế, để dự đoán cơ cấu kinh tế của Việt Nam cho năm 2025 bằng mô hình ARIMA. Xác thực mô hình được hỗ trợ bởi tự động tương quan và phân tích chuỗi thời gian. Dự báo cho giai đoạn 2020 đến 2025 cho thấy một sự thay đổi đáng kể trong bối cảnh kinh tế: đóng góp của ngành công nghiệp và xây dựng dự kiến sẽ giảm xuống còn 38%, trong khi nông, lâm nghiệp và thủy sản được dự đoán sẽ giảm từ 16% năm 2019 xuống còn 12%. Khu vực dịch vụ được dự đoán sẽ tăng từ 46% lên 52% vào năm 2025. Nghiên cứu này có ý nghĩa quan trọng đối với việc xây dựng chính sách kinh tế của Việt Nam".

Đoạn 3: "Nghiên cứu này điều tra ảnh hưởng của các chỉ số tài chính đến hiệu suất của các công ty niêm yết trên thị trường chứng khoán Việt Nam, sử dụng Lợi nhuận trên Tài sản (ROA) làm chỉ số hiệu suất chính. Phân tích dữ liệu từ 621 công ty trong lĩnh vực tiêu dùng và công nghiệp cho năm 2019, sử dụng phân tích phương trình đồng thời và hồi quy với Phương pháp Khoảnh khắc Tổng quát (GMM), xác định ba chỉ số tài chính quan trọng ảnh hưởng đáng kể đến thành công kinh doanh. Yếu tố chính là Lợi nhuận trên Tài sản (ROA) tổng thể, tiếp theo là tỷ lệ nợ trên tài sản và Lợi nhuận trên Vốn chủ sở hữu (ROE). Nghiên cứu cũng lưu ý rằng tỷ lệ nợ trên tài sản có tác động đến ROE. Những phát hiện này thách thức nghiên cứu trước đó của Hiệp hội các Nhà kinh doanh Chứng khoán Việt Nam, cho thấy hiệu quả quản lý kinh doanh phần lớn độc lập với tổng doanh thu tài sản. Các nhà nghiên cứu đề xuất các chiến lược quản lý để tăng hiệu quả hoạt động và cung cấp hướng dẫn cho các nhà đầu tư để đánh giá hiệu suất, tạo điều kiện cho việc tạo ra các danh mục đầu tư được cải thiện".

Đoạn 4: "Một cuộc điều tra tiên phong ở Hawthorne, Hoa Kỳ, đã phát hiện ra rằng trong một nhóm nhân viên thử nghiệm chịu các mức độ ánh sáng khác nhau, việc tăng cường chiếu sáng có liên quan đến hiệu quả cao hơn. Thật bất ngờ, cả hai nhóm thử nghiệm và kiểm soát đều thể hiện một mô hình đầu ra được cải thiện ngay cả dưới ánh sáng nhất quán. Các nhà khoa học nhằm mục đích điều tra các yếu tố hành vi khác nhau liên quan đến những thay đổi kỹ thuật và vật lý để làm rõ những kết quả bất thường này. Mayo và nhóm của ông đã được yêu cầu tham gia nỗ lực nghiên cứu này. Trong suốt 18 tháng, nhóm của Mayo đã cải thiện điều kiện làm việc bằng cách thực hiện thời gian nghỉ theo lịch trình, cung cấp dịch vụ bữa ăn tại nhà và rút ngắn tuần làm việc, dẫn đến năng suất tăng đáng kể. Nghiên cứu tập trung vào các nhân viên nữ lắp ráp rơle điện thoại. Sau đó, môi trường làm việc được khôi phục lại trạng thái ban đầu để dự đoán những tác động tiêu cực đến sản xuất. Tuy nhiên, kết quả cho thấy sản lượng thậm chí còn tăng đáng kể hơn".

Đoạn 5: "Đánh giá ban đầu về hoạt động của công ty chỉ ra rằng hiệu suất hiện tại đang tiếp cận các mục tiêu hoặc dự báo đã đặt ra. Đánh giá toàn diện này cho thấy các lĩnh vực quan trọng cần chú ý trong giai đoạn tiếp theo. Các biến thể xảy ra do ảnh hưởng trực tiếp của các yếu tố tác động đến các chỉ số hiệu suất của công ty. Xác định và đánh giá các yếu tố này dẫn đến việc phát hiện ra các cơ hội tăng trưởng, các lĩnh vực cần nâng cao và các điểm yếu nội bộ cần giải quyết. Bằng cách nhận ra những tiến bộ trong việc thực hiện và sự khác biệt so với dự đoán ban đầu, các doanh nghiệp có thể sửa đổi kế hoạch tương lai để phù hợp với mục tiêu tăng trưởng. Thủ tục đánh giá này rất cần thiết cho các tổ chức nhằm xác định chính xác các sửa đổi cần thiết cho giai đoạn sắp tới".

3.2. Phương pháp phân tích

Đầu ra của thuật toán phát hiện AI đã được xử lý tỉ mỉ để loại bỏ dữ liệu không đầy đủ hoặc không chính xác trước khi phân tích. Mỗi mẫu văn bản được phân tích để xác định tỷ lệ nội dung do con người và AI tạo ra. Các thước đo thống kê mô tả, chẳng hạn như mức trung bình và tỷ lệ phần trăm, được sử dụng để đánh giá hiệu quả của chương trình. Hiệu quả của các hệ thống phát hiện AI đã được đánh giá thông qua phân tích đường cong ROC. Jamovi phiên bản 2.6.2 đã được sử dụng để phân tích dữ liệu. Nghiên cứu này đã áp dụng phương pháp đường cong ROC để đánh giá khả năng của phần mềm phát hiện AI trong việc phân biệt giữa văn bản của con người và AI. Đầu ra của chương trình nhận dạng AI đóng vai trò là các biến phụ thuộc, trong khi các mẫu văn bản khác nhau tạo thành các biến độc lập. Nghiên cứu đã điều tra tiền đề rằng chương trình phát hiện AI thể hiện khả năng phân biệt cao, được biểu thị bằng giá trị AUC vượt quá 0.7.

4. Kết quả nghiên cứu và thảo luận

4.1. Thống kê mô tả

Bảng 1

Thống Kê Mô Tả

	HM	**AI**	**AV**
N	150	150	150
Missing	0	0	0
Mean	49.7	42.7	46.2
Std. error mean	3.12	3.08	1.01
Median	50.8	32.8	48.5
Standard deviation	38.2	37.8	12.4
Minimum	0	0	0.5
Maximum	100	100	77
20th percentile	2.89	2.38	43.2
40th percentile	32.6	21.9	47.1
60th percentile	68	59.8	49.8
80th percentile	95.2	90.7	51.5

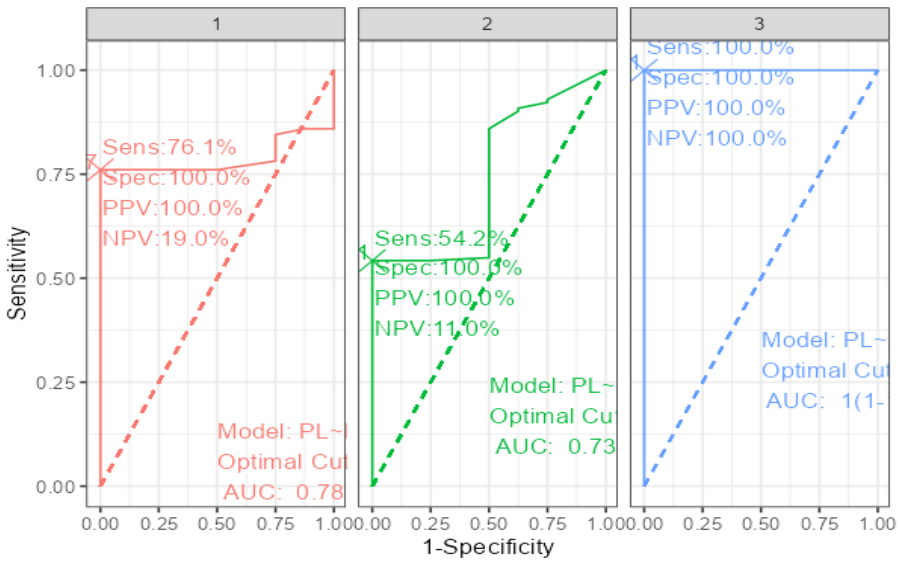
Nguồn: Kết quả phân tích của các tác giả

Bảng 1 hiển thị các số liệu thống kê khác nhau cho ba biến (HM, AI, AV) từ một mẫu gồm 150 người tham gia. Các đặc điểm phân bố cho thấy HM đối xứng, AI nghiêng phải và AV hơi lệch sang trái. HM và AI thể hiện sự thay đổi đáng kể với độ lệch chuẩn cao (38.2 và 37.8), trong khi AV thể hiện sự phân bố tập trung hơn với độ lệch chuẩn thấp hơn (12.4). Phạm vi của AV là 0.5 đến 77.0, trong khi HM và AI trải dài từ 0 đến 100. Mặc dù phạm vi hẹp hơn của AV, HM và AI thể hiện sự thay đổi phần trăm đáng kể, đặc biệt là ở phần trăm thứ 60 và 80. Khả năng phân biệt giữa AI và văn bản do con người tạo ra của phần mềm đã cho thấy sự cải thiện dần dần. Độ lệch chuẩn lớn trong HM và AI phản ánh khả năng phát hiện đa dạng, có thể là kết quả của các thuật toán hoặc chất lượng văn bản khác nhau. Sự khác biệt giữa giá trị trung bình và giá trị trung bình, đặc biệt là đối với AI, làm nổi bật bản chất không đối xứng của dữ liệu, đảm bảo sự thận trọng trong việc áp dụng các kỹ thuật thống kê tham số. Sự nhất quán được nâng cao trong các đánh giá AV cho thấy rằng việc tích hợp các phương pháp khác nhau có thể tăng cường độ tin cậy. Để nâng cao độ chính xác, điểm trung bình cho HM và AI giảm xuống dưới 50%. Các biến thể ở phần trăm thứ 60 và 80 biểu thị sự tồn tại của nhiều văn bản hoặc ứng dụng với kết quả đặc biệt.

4.2. Phân tích đường cong ROC

Hình 1

Đường Cong ROC



Nguồn: Kết quả phân tích của các tác giả

Đường cong ROC cho HM, AI và AV cho thấy sức mạnh phân biệt vượt trội so với cơ hội. Phân tích thống kê cho thấy giá trị AUC của $p = 0.001$ cho cả HM và AV, và $p = 0.001$ cho AI. AUC của HM là 0.787, AI là 0.738 và AV là 1.0, thể hiện hiệu suất mạnh mẽ cho HM và AI, với AV đạt được sự phân loại hoàn hảo. Điều này cho thấy khả năng phân biệt mạnh mẽ trong các hệ thống phát hiện văn bản AI, đặc biệt là trong hiệu suất trung bình (AV). Sự khác biệt AUC là 0.0493 giữa HM và AI là không đáng kể ($p = 0.558$). Sự khác biệt đáng kể về AUC là -0.2130 được quan sát thấy giữa HM và AV ($p < 0.001$). Sự khác biệt AUC là -0.2623 giữa AI và AV cũng đáng kể ($p < 0.001$). Nhận dạng văn bản của con người và AI thể hiện hiệu suất tương tự, nhưng AV vượt qua cả hai. Giá trị Z là 44.6 ($p < 0.001$) làm nổi bật những thay đổi đáng kể trong đường cong ROC, cho thấy sự thay đổi về hiệu quả phân loại giữa các phương pháp phát hiện.

Phân tích đường cong ROC chỉ ra rằng các công cụ nhận dạng văn bản AI sở hữu khả năng phân biệt đáng chú ý, với giá trị AUC vượt quá 0.7 cho cả nội dung do con người và AI tạo ra. Điều này nhấn mạnh tầm quan trọng của các công cụ này trong việc phân biệt văn bản do máy sản xuất với văn bản do con người tạo ra. Hệ thống phát hiện đã chứng minh hiệu quả nhất quán, cho thấy không có sự chênh lệch đáng kể nào trong việc xác định nội dung do con người và AI tạo ra. AUC trung bình được tính là 1 khi tích hợp phương pháp biểu thị hiệu suất tối ưu, minh họa rằng các kỹ thuật tính trung bình nâng cao độ chính xác trên các phương pháp phát hiện khác nhau. Mặc dù có khả năng đáng khen ngợi trong việc phân biệt con người với văn bản AI, các giá trị AUC dưới 0.8 cho thấy sự cần thiết phải cải tiến và điều tra thêm về các khuôn khổ phát hiện văn bản AI. Tính nhất quán của các phát hiện được chứng thực bởi các giá trị p thấp (dưới 0.001) trong các so sánh khác nhau. Phân tích ROC, đặc biệt là trong bối cảnh của các phương pháp phát hiện kết hợp, cho thấy những tác động thực tế đáng kể đối với các công nghệ nhận dạng văn bản AI.

4.3. Thảo luận kết quả phân tích

Nghiên cứu này sử dụng phân tích đường cong ROC và thống kê mô tả để đánh giá hiệu quả của hệ thống nhận dạng văn bản AI. Thuật toán đạt được hiệu suất vượt trội (AUC = 1) trong việc phân biệt giữa các văn bản do AI tạo ra và do con người tạo ra. Mặc dù thiếu ý nghĩa thống kê ($p = 0.558$), các phát hiện mô tả cho thấy độ chính xác nhận dạng vượt trội đối với nội dung do con người tạo ra (trung bình = 49.7) so với các văn bản do AI tạo ra (trung bình = 42.7). Kết quả này hỗ trợ các mục tiêu phát hiện văn bản AI đương đại, phản ánh khả năng phát hiện ấn tượng của máy tính cho cả hai loại văn bản. Sự khác biệt về hiệu quả phát hiện nhỏ đã được ghi nhận cho cả hai loại văn bản - HM, độ lệch chuẩn = 38.2; AI, độ lệch chuẩn = 3.8 - có thể được quy cho sự thay đổi về chất lượng văn bản hoặc hiệu suất phần mềm. Phát hiện này được củng cố bởi Weber-Wulff và cộng sự (2023), Legaspi và cộng sự (2024), những người cũng thảo luận về sự không nhất quán về hiệu suất trong nhận dạng văn bản AI. Mặc dù tiềm năng sáng tạo vốn có, phân tích ROC cho thấy khả năng phân biệt đáng kể cho cả HM và AI (AUC > 0.7). Cái nhìn sâu sắc này nhấn mạnh những tiến bộ nhanh chóng trong công nghệ AI, đòi hỏi phải có những cải tiến liên tục trong các phương pháp phát hiện để duy trì hiệu quả. Kết luận này cộng hưởng với quan điểm của Miller và Lin (2022), Alshahrani (2024), nhấn mạnh sự cần thiết của những tiến bộ bền vững trong lĩnh vực này.

Nghiên cứu tiết lộ rằng phương pháp kết hợp thể hiện hiệu suất tối ưu. Điều này chỉ ra rằng việc tích hợp các mức trung bình với các thuật toán đa dạng giúp tăng cường độ chính xác. Các phát hiện nhấn mạnh tiềm năng phát triển các hệ thống phát hiện tiên tiến trong bối cảnh tương lai. Tuy nhiên, nghiên cứu này có những hạn chế đáng chú ý cần được xem xét. Kích thước mẫu 150 phiên bản cho mỗi danh mục có thể không bao gồm toàn bộ nội dung xác thực. Ngoài ra, việc thiếu sự khác biệt giữa các thể loại văn bản có thể ảnh hưởng đến hiệu quả phát hiện.

Các nghiên cứu tiếp theo sẽ tăng số lượng người tham gia và mở rộng phân tích văn bản để giảm thiểu những hạn chế này và đạt được những hiểu biết toàn diện hơn. Một cuộc điều tra kỹ lưỡng về các yếu tố ảnh hưởng đến hiệu quả phát hiện, chẳng hạn như kích thước văn bản và phong cách viết, có thể mang lại những phát hiện đáng kể. Nghiên cứu này đã cung cấp bằng chứng đáng kể xác nhận hiệu quả của các công cụ phát hiện văn bản AI, đặc biệt là thông qua cách tiếp cận toàn diện. Các kết quả nâng cao việc đánh giá các công nghệ hiện có và mở đường cho những tiến bộ trong phương pháp phát hiện văn bản AI trong tương lai.

5. Kết luận

Nghiên cứu này sử dụng thống kê mô tả và phân tích đường cong ROC để đánh giá hiệu quả của công cụ phát hiện văn bản AI. Phần mềm đã thể hiện hiệu suất vượt trội (AUC = 1.), phân biệt chính xác giữa văn bản do AI tạo ra và văn bản do con người tạo ra. Những phát hiện này phù hợp với Pandita và cộng sự (2024). Thống kê mô tả cho thấy độ chính xác cao hơn trong việc xác định văn bản do con người tạo ra (trung bình HM = 49.7) so với nội dung do AI tạo ra (AI trung bình = 42.7). Tuy nhiên, phân tích ROC cho thấy sự khác biệt này thiếu ý nghĩa thống kê (p = 0.558). Điều này ngụ ý độ tin cậy của phần mềm trong việc phân biệt cả hai loại văn bản, phù hợp với các mục tiêu nhận dạng văn bản AI đương đại và tương phản với các nghiên cứu trước đó (Rashidi & ctg., 2023; Walters, 2023; Weber-Wulff & ctg., 2023). Hiệu quả vượt trội của phương pháp (AUC = 1) nhấn mạnh tầm quan trọng của việc tích hợp nhiều thuật toán phát hiện để nâng cao độ chính xác. Kết luận này chứng thực nghiên cứu của Cui và cộng sự (2023). Những kết quả này nhấn mạnh tiềm năng tiến bộ trong các khung phát hiện tích hợp, sở hữu những phân nhánh đáng kể trong thế giới thực.

Quan điểm và hướng nghiên cứu trong tương lai

Nghiên cứu này trình bày những phát hiện quan trọng về hiệu quả của hệ thống phát hiện văn bản do AI tạo ra. Các cải tiến có thể cải thiện kết quả và mở rộng khả năng áp dụng:

Mở rộng quy mô mẫu và sự đa dạng: Sử dụng các nhóm thí nghiệm lớn hơn và đa dạng hơn sẽ tăng độ tin cậy và mức độ liên quan của nghiên cứu. Kết hợp các định dạng văn bản đa dạng, bao gồm các bài báo học thuật, thông cáo báo chí, hồ sơ chính thức và bài đăng trên phương tiện truyền thông xã hội, sẽ mang lại một đánh giá toàn diện về hiệu quả của hệ thống. Điều tra các yếu tố ảnh hưởng: Cần nghiên cứu thêm để xác định tác động của độ dài văn bản đến khả năng phát hiện của hệ thống, lưu ý rằng các đoạn trích ngắn hơn được xác định dễ dàng hơn. Ngoài ra, điều quan trọng là phải kiểm tra chủ đề và phong cách viết ảnh hưởng đến khả năng phát hiện ra như thế nào, vì các đặc điểm ngôn ngữ khác nhau có thể làm cho các chủ đề hoặc phong cách nhất định dễ nhận biết hơn.

Phương pháp phát hiện nâng cao: Các kỹ thuật học tập nâng cao phải được nghiên cứu để tăng cường nhận dạng văn bản AI, đặc biệt là liên quan đến các chiến lược trốn tránh. Điều quan trọng là phải tích hợp các phương pháp tìm kiếm ngữ nghĩa để tăng cường độ chính xác, ngay cả khi tài liệu bị thay đổi.

Phát triển khung phát hiện toàn diện: Tăng cường độ chính xác có thể được thực hiện thông qua việc phân tích các kết quả trung bình và kết hợp các kỹ thuật nhận dạng khác nhau. Việc tổng hợp các phương pháp đa dạng, được minh họa bằng phương pháp kết hợp (AV), mang lại kết quả vượt trội, được chứng minh bằng điểm hiệu suất tối ưu (AUC = 1). Cập nhật liên tục các hệ thống phát hiện là rất quan trọng để chống lại các chiến lược trốn tránh mới nổi và tiến bộ trong AI.

Đánh giá hiệu quả thực tế: Đánh giá hiệu quả của chương trình là rất quan trọng đối với khả năng ứng dụng của nó trên các lĩnh vực truyền thông, thương mại và giáo dục. Thu thập phản hồi của người dùng là công cụ trong việc xác định các thách thức thực hiện và thúc đẩy các giải pháp sáng tạo. Việc áp dụng các phương pháp tiên tiến này không chỉ cải thiện độ tin cậy và mức độ liên quan của các hệ thống phát hiện văn bản AI mà còn bảo vệ tính toàn vẹn và minh bạch của thông tin trong bối cảnh kỹ thuật số đương đại.

Tài liệu tham khảo

Akram, A. (2023). *An empirical study of AI generated text detection tools.* https://doi.org/10.33140/AMLAI

Alshahrani, A. (2024). Artificial intelligence technologies utilization for detecting explosive materials. *International Journal of Data and Network Science, 8*(1), 617-628.

Bellini, V., Semeraro, F., Montomoli, J., Cascella, M., & Bignami, E. (2024). Between human and AI: Assessing the reliability of AI text detection tools. *Current Medical Research and Opinion, 40*(3), 353-358. https://doi.org/10.1080/03007995.2024.2310086

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). *On the possibilities of AI-generated text detection*. https://doi.org/10.48550/arXiv.2304.04736

Cui, W., Zhang, L., Wang, Q., & Cai, S. (2023). *Who said that? Benchmarking social media AI detection*. http://arxiv.org/abs/2310.08240

Hu, X., Chen, P. Y., & Ho, T. Y. (2023). *RADAR: Robust AI-text detection via adversarial learning.* https://doi.org/10.48550/arXiv.2307.03838

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). *Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense.* https://doi.org/10.48550/arXiv.2303.13408

Kumarage, T., Sheth, P., Moraffah, R., Garland, J., & Liu, H. (2023). *How reliable are AI-generated-text detectors? An assessment framework using evasive soft prompts.* https://doi.org/10.48550/arXiv.2310.05095

Legaspi, J. B., Licuben, R. J. O., Legaspi, E. A., & Tolentino, J. A. (2024). Comparing AI detectors: Evaluating performance and efficiency. *International journal of Science and Research Archive, 12*(2), 833-838.

Lu, N., Liu, S., He, R., & Tang, K. (2023). *Large language models can be guided to evade AI-generated text detection.* https://doi.org/10.48550/arXiv.2305.10847

Miller, I., & Lin, D. (2022). *Developing self-evolving deepfake detectors against AI attacks* [Paper presentation]. 2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), Atlanta, GA. https://ieeexplore.ieee.org/abstract/document/10063474/

Pandita, V., Mujawar, A. M., Norbu, T., Verma, V., & Patil, P. (2024). *Text origin detection: Unmasking the source - AI vs human* [Paper presentation]. 2024 5th International Conference for Emerging Technology (INCET), Belgaum, India. https://ieeexplore.ieee.org/abstract/document/10592990/

Rashidi, H. H., Fennell, B. D., Albahra, S., Hu, B., & Gorbett, T. (2023). The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Journal of Pathology Informatics, 14*, Article 100342.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected?* https://doi.org/10.48550/arXiv.2303.11156

Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2023). *Ghostbuster: Detecting text ghostwritten by large language models*. https://doi.org/10.48550/arXiv.2305.15047

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, *7*. https://doi.org/10.1515/opis-2022-0158

Wang, Y. (2024). Survey for detecting AI-generated content. *Advances in Engineering Technology Research*, *11*(1), 643-643.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), Article 26. https://doi.org/10.1007/s40979-023-00146-z