GENERATIVE ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF FAKE NEWS DETECTION IN DIGITAL MEDIA

TRÍ TUỆ NHÂN TẠO TẠO SINH VÀ BÀI TOÁN PHÁT HIỆN TIN GIẢ TRONG TRUYỀN THÔNG SỐ

Trần Văn Cường
Trường Đại học Quảng Bình

ABSTRACT: *The rapid advancement of Generative Artificial Intelligence (GenAI) has significantly increased the sophistication, scale, and dissemination speed of fake news in the contemporary digital media environment. Large language models, deepfake technologies, and multimodal generative systems enable the creation of highly realistic fabricated content, posing substantial challenges to fact-checking and verification processes. This paper examines the characteristics of fake news in the era of Artificial Intelligence (AI), reviews current detection approaches, and highlights the dual role of AI as both an effective tool for information verification and a critical challenge due to its evolving capability to evade detection mechanisms. Furthermore, the dynamic and adversarial nature of fake news detection in the AI era is analyzed. Based on this analysis, the study proposes an integrated, multimodal, and adaptive framework to enhance the effectiveness and robustness of fake news detection.*

Keywords: *Generative artificial intelligence, fake news detection, information verification.*

TÓM TẮT: *Sự phát triển của trí tuệ nhân tạo tạo sinh (Generative Artificial Intelligence - GenAI) đã làm tăng mức độ tinh vi, quy mô và tốc độ lan truyền của tin giả trong môi trường truyền thông số hiện nay. Các mô hình ngôn ngữ lớn, công nghệ deepfake và hệ thống tạo sinh đa phương thức có thể cho phép tạo nội dung giả có độ chân thực cao, gây khó khăn cho hoạt động kiểm chứng. Bài báo này phân tích đặc trưng của tin giả trong kỷ nguyên trí tuệ nhân tạo (Artificial Intelligence - AI), tổng hợp các phương pháp phát hiện hiện nay, đồng thời chỉ ra tính hai mặt của AI: vừa là công cụ hỗ trợ kiểm chứng, vừa là thách thức do khả năng cải thiện và né tránh phát hiện. Bản chất động và đối kháng của bài toán phát hiện tin giả trong kỷ nguyên AI cũng được làm rõ, từ đó đề xuất định hướng tiếp cận tích hợp, đa phương thức và có khả năng thích ứng cao trong phát hiện tin giả.*

Từ khóa*: Trí tuệ nhân tạo tạo sinh, phát hiện tin giả, kiểm chứng thông tin.*

1. ĐẶT VẤN ĐỀ

Trong kỷ nguyên chuyển đổi số, thông tin trở thành nguồn lực chiến lược, là yếu tố quyết định trong đời sống chính trị, kinh tế, văn hóa - xã hội. Tuy nhiên, với sự bùng nổ của truyền thông số là sự gia tăng nhanh chóng của tin giả (fake news), thông tin sai lệch (misinformation) và thông tin có chủ đích thao túng (disinformation) [1].

Ngày nay, với sự hỗ trợ của GenAI, việc tạo ra văn bản, hình ảnh, video hoặc giọng nói giả với độ chân thực cao trở nên đơn giản. Công nghệ deepfake có thể tái tạo khuôn mặt và giọng nói của người nổi tiếng; các mô hình ngôn ngữ lớn có thể tạo ra bài viết, bình luận hoặc "bằng chứng" giả

mang tính thuyết phục cao [2]. Các hệ thống tạo ảnh AI có thể sinh ra hình ảnh về những sự kiện chưa từng xảy ra như thật. Thực tế đã xuất hiện nhiều vụ việc lan truyền hình ảnh giả về thảm họa, phát biểu giả mạo của lãnh đạo hay video cắt ghép nhằm thao túng dư luận trên mạng xã hội [8]. Các nội dung này thường được lan truyền với tốc độ nhanh hơn so với quá trình kiểm chứng và gây ra các hệ quả xấu.

Bài toán phát hiện tin giả không còn đơn thuần là vấn đề kỹ thuật xử lý văn bản hay phân tích hình ảnh mà đã trở thành một thách thức liên ngành. Nó kết hợp giữa khoa học dữ liệu, truyền thông, an ninh mạng và chính sách công. Đặc biệt khi mà GenAI có khả năng liên tục tự cải thiện và tối ưu hóa nội dung đầu ra thì các phương pháp phát hiện truyền thống chỉ dựa trên dấu hiệu bề mặt trở nên kém hiệu quả [2]. Việc nghiên cứu một cách hệ thống về tin giả trong kỷ nguyên GenAI, đồng thời phân tích các phương pháp phát hiện và những thách thức mới đặt ra có ý nghĩa thực tiễn cao. Đây không chỉ là vấn đề công nghệ mà còn là yêu cầu cấp thiết nhằm bảo vệ tính minh bạch, độ tin cậy và an toàn của không gian thông tin số. Mục tiêu của bài báo này là phân tích tác động của GenAI đối với sự biến đổi của tin giả trong môi trường truyền thông số, đồng thời hệ thống hóa các hướng tiếp cận phát hiện tin giả hiện nay dưới góc nhìn công nghệ. Trên cơ sở đó, bài báo làm rõ các thách thức mới phát sinh trong quá trình kiểm chứng thông tin khi các mô hình tạo sinh ngày càng phát triển mạnh. Qua đó, nghiên cứu góp phần cung cấp một khung nhìn tổng hợp về bài toán phát hiện tin giả trong kỷ nguyên AI và gợi mở các hướng tiếp cận tích hợp, đa tầng nhằm nâng cao hiệu quả kiểm chứng thông tin trong môi trường truyền thông số.

2. TRÍ TUỆ NHÂN TẠO TẠO SINH VÀ SỰ BIẾN ĐỔI CỦA TIN GIẢ TRONG TRUYỀN THÔNG SỐ

2.1. Khái niệm và phân loại GenAI

GenAI là một nhánh của AI, nó tập trung vào học phân phối xác suất của dữ liệu huấn luyện nhằm tạo ra dữ liệu mới có đặc trưng tương đồng. Khác với các hệ thống AI truyền thống chủ yếu thực hiện các nhiệm vụ như phân loại, dự đoán, ra quyết định,… thì GenAI có khả năng sinh ra nội dung mới dưới nhiều dạng khác nhau. Dựa trên loại dữ liệu đầu ra, GenAI có thể được phân loại thành [4]:

- Mô hình tạo sinh văn bản: là các mô hình có khả năng sinh chuỗi ngôn ngữ tự nhiên theo ngữ cảnh đầu vào. Tiêu biểu là các mô hình ngôn ngữ lớn (Large Language Models - LLMs), thường được xây dựng dựa trên kiến trúc Transformer.

Mô hình tạo sinh hình ảnh: là các mô hình tổng hợp dữ liệu thị giác dưới dạng ảnh mới từ dữ liệu huấn luyện hoặc tín hiệu điều kiện hóa. Các phương pháp phổ biến gồm mạng đối nghịch tạo sinh (Generative Adversarial Networks - GAN) và mô hình khuếch tán (Diffusion Models) [3].

- Mô hình tạo sinh âm thanh và giọng nói: là các mô hình tạo tín hiệu âm thanh theo miền thời gian, bao gồm tổng hợp giọng nói và sinh âm thanh tổng hợp, dựa trên đặc trưng học được từ các dữ liệu đã được huấn luyện.
- Sinh mã lập trình: nhóm mô hình này tạo ra hoặc biến đổi chương trình máy tính dưới dạng mã nguồn. Hỗ trợ các tác vụ như sinh hàm, hoàn thiện mã, chuyển đổi ngôn ngữ lập trình, gợi ý sửa lỗi,…

- Mô hình đa phương thức: là các hệ thống có khả năng xử lý và sinh đầu ra là kết hợp của nhiều dạng dữ liệu khác nhau như văn bản, hình ảnh, âm thanh,… vào trong một kiến trúc thống nhất [5].

2.2. Cơ chế tạo sinh nội dung và tác động đối với thông tin sai lệch

Xét về nguyên lý hoạt động, các GenAI không thực hiện đánh giá giá trị đúng hay sai của thông tin mà nó sinh ra. Nó tối ưu hóa xác suất xuất hiện của chuỗi dữ liệu phù hợp với ngữ cảnh đầu vào dựa trên phân phối xác suất học được từ dữ liệu huấn luyện. Đặc điểm này đã dẫn đến một số hệ quả đáng chú ý đối với hiện tượng thông tin sai lệch do các GenAI tạo ra.

Thứ nhất, các mô hình có khả năng tạo ra nội dung có cấu trúc logic, ngôn ngữ mạch lạc và có phong cách diễn đạt gần tương đồng với các nguồn thông tin chính thống. Do đó có thể làm gia tăng mức độ thuyết phục cảm nhận của thông tin được tạo sinh.

Thứ hai, quá trình tạo sinh có thể được điều kiện hóa thông qua yêu cầu đầu vào, cho phép người sử dụng định hướng nội dung theo mục đích cụ thể (bao gồm cả việc tạo lập và khuếch đại thông tin sai lệch).

Một vấn đề quan trọng liên quan đến nội dung do GenAI tạo ra là hiện tượng ảo giác mô hình (hallucination). Trong trường hợp này, mô hình có thể tạo ra các thông tin có vẻ hợp lý về mặt ngôn ngữ nhưng không có cơ sở dữ liệu thực tế hoặc không tồn tại trong các nguồn dữ liệu đáng tin cậy.

2.3. Đặc trưng nội dung do GenAI tạo ra dưới góc nhìn phát hiện

Nội dung do các mô hình GenAI tạo ra có một số đặc trưng về thống kê, ngôn ngữ và cấu trúc biểu đạt. Các đặc trưng này thường được khai thác trong các phương pháp phát hiện để phân biệt giữa văn bản do con người viết và văn bản được tạo ra bởi GenAI.

2.3.1. Đặc trưng thống kê (Statistical features)

Văn bản do các mô hình ngôn ngữ tạo ra thường có phân bố xác suất từ vựng ổn định theo ngữ cảnh. Do các mô hình GenAI được huấn luyện với mục tiêu dự đoán chuỗi từ có xác suất cao nhất, những từ hoặc cụm từ phổ biến trong ngữ cảnh hiện tại có xu hướng được lựa chọn cao [9]. Vì vậy, văn bản do AI tạo ra thường có hỗn loạn (perplexity) thấp hơn và mức độ dự đoán cao hơn so với văn bản do con người viết. Ngoài ra, sự biến động tự nhiên trong phân bố từ vựng của văn bản tạo sinh (burstiness) thấp do quá trình sinh nội dung ưu tiên các cấu trúc ngôn ngữ phổ biến và ổn định.

2.3.2. Đặc trưng ngôn ngữ (Linguistic features)

Nội dung do AI tạo ra thường có mức độ trôi chảy ngữ pháp cao và tỷ lệ lỗi hình thức thấp. Các mô hình ngôn ngữ lớn được huấn luyện trên khối lượng dữ liệu văn bản lớn nên có khả năng tạo ra các câu hoàn chỉnh, đúng cấu trúc ngữ pháp và phù hợp với ngữ cảnh [9]. Tuy nhiên, sự trôi chảy này đôi khi dẫn đến hiện tượng "chuẩn hóa quá mức", trong đó các câu văn có cấu trúc tương đối đồng đều và ít xuất hiện các biến thể ngôn ngữ mang tính cá nhân hoặc ngẫu nhiên như trong văn bản của con người.

2.3.3. Đặc trưng cấu trúc (Structural patterns)

Bên cạnh các đặc trưng thống kê và ngôn ngữ, văn bản do AI tạo ra cũng có thể biểu hiện một số mẫu cấu trúc lặp lại trong cách tổ chức nội dung. Các đoạn văn thường được xây dựng theo những khuôn

mẫu diễn đạt quen thuộc, chẳng hạn như cấu trúc giải thích theo từng bước, liệt kê có trật tự hoặc các mẫu lập luận mang tính hệ thống [10], [11]. Những đặc điểm này góp phần tạo nên tính mạch lạc của văn bản.

Các đặc trưng trên cung cấp những tín hiệu quan trọng cho các phương pháp phát hiện nội dung tạo sinh. Tuy nhiên, khi các mô hình GenAI ngày càng được cải tiến thì ranh giới giữa văn bản của con người và văn bản do AI tạo ra ngày càng trở nên mờ nhạt. Nó đặt ra nhiều thách thức mới cho các hệ thống phát hiện văn bản do AI tạo ra.

3. THỰC TIỄN TIN GIẢ ĐƯỢC TẠO RA TỪ AI

Sự phát triển của các hệ GenAI không chỉ làm thay đổi phương thức sản xuất nội dung trên phương diện lý thuyết mà còn tạo ra những biểu hiện tác động trực tiếp trong các lĩnh vực chính trị, kinh tế và xã hội. Các hiện tượng như video deepfake, giả mạo hình ảnh và tổng hợp giọng nói cho thấy thông tin sai lệch trong kỷ nguyên AI đã có ảnh hưởng đến nhận thức công chúng và môi trường truyền thông.

3.1. Các hình thức tin giả

Sự phát triển của GenAI đã làm gia tăng các hình thức tin giả dựa trên khả năng giả mạo danh tính với mức độ chân thực cao. Các công nghệ như deepfake, tổng hợp giọng nói và mô hình ngôn ngữ lớn cho phép tạo nội dung mô phỏng con người, từ đó tác động trực tiếp đến nhận thức và hành vi của công chúng.

Trong lĩnh vực chính trị, deepfake được sử dụng để tạo video giả mạo phát biểu của lãnh đạo. Điển hình là video giả Tổng thống Ukraine Volodymyr Zelenskyy kêu gọi đầu hàng (2022) hay đoạn ghi âm giả liên quan đến Michal Simecka trong bầu cử Slovakia (2023). Các nội dung này lan truyền nhanh chóng trước khi được kiểm chứng. Làm gia tăng nguy cơ thao túng dư luận và gây nhiễu loạn thông tin.

Trong truyền thông và thương mại, GenAI cho phép giả mạo hình ảnh và giọng nói người nổi tiếng trong quảng cáo hoặc lừa đảo, khai thác uy tín cá nhân để tăng độ tin cậy cảm nhận. Đồng thời, khả năng tạo nội dung cá nhân hóa ở quy mô lớn làm mờ ranh giới giữa ứng dụng hợp pháp và giả mạo.

Trong lĩnh vực tài chính, các hình thức lừa đảo ngày càng tinh vi với việc giả mạo giọng nói lãnh đạo để yêu cầu giao dịch khẩn cấp, kết hợp chatbot có khả năng hội thoại đa lượt và thích ứng ngữ cảnh. Việc tích hợp dữ liệu cá nhân và kỹ thuật thao túng xã hội làm tăng hiệu quả lừa đảo và gây khó khăn cho các phương pháp phát hiện truyền thống.

Tin giả dựa trên giả mạo danh tính do GenAI tạo ra có tính đa dạng, quy mô lớn và mức độ tinh vi cao, đặt ra yêu cầu phát triển các phương pháp phát hiện và kiểm chứng phù hợp trong môi trường truyền thông số.

3.2. Đặc điểm mới của tin giả thời AI

Có thể khái quát một số đặc điểm cấu thành của thông tin sai lệch trong bối cảnh GenAI như sau:

Tính đa phương thức

Thông tin sai lệch không còn chỉ giới hạn ở văn bản mà được thể hiện dưới dạng kết hợp nhiều phương thức dữ liệu như hình ảnh, âm thanh và video. Sự tích hợp này làm tăng mức độ cảm nhận và nâng cao hiệu quả thuyết phục đối với người tiếp nhận.

Khả năng mở rộng quy mô

Quá trình tạo lập nội dung được tự động hóa ở mức cao, cho phép triển khai và duy trì hoạt động phát tán thông tin sai lệch

trên quy mô lớn trong thời gian ngắn. Ví dụ, các hệ thống tạo nội dung dựa trên mô hình GenAI có thể tạo ra số lượng lớn bài viết, bình luận hoặc bài đăng mạng xã hội chỉ trong thời gian ngắn. Khi kết hợp với mạng bot hoặc tài khoản tự động thì các nội dung này có thể được phát tán đồng thời trên nhiều nền tảng khác nhau làm gia tăng tốc độ lan truyền của thông tin sai lệch.

Giảm chi phí sản xuất

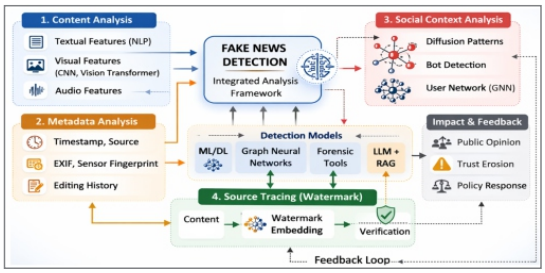
GenAI làm giảm đáng kể yêu cầu về nguồn lực kỹ thuật và chi phí cần thiết để tạo nội dung giả mạo có chất lượng cao. Việc giảm rào cản công nghệ có thể góp phần mở rộng đối tượng có khả năng tham gia tạo ra nội dung sai lệch.

Gia tăng khó khăn trong phát hiện bằng cảm quan

Chất lượng biểu đạt của thông tin ngày càng tinh vi khiến việc phân biệt nội dung thật - giả bằng trực giác trở khó khăn. Làm suy giảm vai trò của các cơ chế kiểm chứng truyền thống và đặt ra yêu cầu phát triển các phương pháp phát hiện dựa trên học máy, phân tích thống kê và khung đánh giá đa phương thức.

4. CÁC PHƯƠNG PHÁP PHÁT HIỆN TIN GIẢ TRONG KỶ NGUYÊN AI

Trong bối cảnh GenAI ngày càng phát triển mạnh, phát hiện tin giả đã trở thành một bài toán liên ngành, kết hợp xử lý ngôn ngữ tự nhiên, thị giác máy tính, phân tích mạng xã hội và kỹ thuật truy vết nguồn gốc. Các phương pháp phát hiện tin giả hiện nay có thể được chia thành bốn nhóm chính (Hình 1) gồm: dựa trên nội dung (Content-based), dựa trên siêu dữ liệu (Metadata-based), dựa trên ngữ cảnh mạng xã hội (Social-context) và dựa trên cơ chế watermark - truy xuất nguồn gốc.



Hình 1. Khung phát hiện tin giả đa tầng trong kỷ nguyên AI tạo sinh

4.1. Phát hiện dựa trên nội dung

Phương pháp dựa trên nội dung tập trung vào phân tích trực tiếp dữ liệu văn bản, hình ảnh, âm thanh hoặc video nhằm phát hiện dấu hiệu bất thường trong nội dung. Cách tiếp cận này giả định rằng nội dung do AI tạo ra mang các đặc trưng thống kê hoặc cấu trúc khác biệt so với dữ liệu tự nhiên do con người tạo ra [6], [7].

Phân tích đặc trưng ngôn ngữ

Các đặc trưng ngôn ngữ như phân bố từ vựng, độ phức tạp cú pháp, mức độ lặp lại và xác suất chuỗi token được sử dụng làm tín hiệu nhận diện. Các mô hình dựa trên perplexity hoặc phân tích xác suất token có thể khai thác sự khác biệt thống kê giữa văn bản do con người và văn bản do mô hình ngôn ngữ lớn tạo ra. Tuy nhiên, khi mô hình sinh ngày càng tiên tiến, sự khác biệt về mặt ngôn ngữ có xu hướng thu hẹp, làm giảm hiệu quả của cách tiếp cận này.

Phân tích hình ảnh, video

Trong deepfake, các phương pháp phát hiện tập trung vào việc xác định bất nhất trong chuyển động, ánh sáng, đổ bóng, tần số nhiễu hoặc đặc trưng sinh học (ví dụ như tần suất chớp mắt bất thường). Những tín hiệu này phản ánh hạn chế trong quá trình tổng hợp dữ liệu thị giác. Tuy nhiên, sự phát triển của các mô hình khuếch tán và kỹ thuật hậu xử lý ngày càng làm suy giảm

các dấu hiệu bất thường dễ quan sát. Tuy nhiên, với sự phát triển của mô hình khuếch tán, chất lượng hình ảnh ngày càng cao làm giảm các dấu hiệu giả mạo rõ ràng [3].

Mô hình deep learning và phát hiện đa phương thức

Các kiến trúc học sâu được sử dụng để học biểu diễn đặc trưng trực tiếp từ dữ liệu, cho phép tích hợp tín hiệu từ nhiều phương thức như văn bản, hình ảnh và âm thanh trong một khung phát hiện thống nhất. Xu hướng hiện nay chuyển từ phát hiện đơn phương thức sang phát hiện đa phương thức nhằm khai thác tính bổ trợ giữa các nguồn tín hiệu.

4.2. Phát hiện dựa trên siêu dữ liệu

Phát hiện dựa trên siêu dữ liệu khai thác các dấu vết ngoại sinh gắn với quá trình tạo lập, xử lý và phân phối nội dung số. Giả định rằng nội dung do GenAI sinh ra thường có các bất thường trong thông tin phụ trợ hoặc trong chu trình xử lý dữ liệu [12].

Phân tích mô hình tạo và phân phối nội dung

Các đặc trưng liên quan đến thời gian tạo, tần suất xuất hiện và mô hình hoạt động có thể cung cấp các tín hiệu nhận diện bất thường. Sự xuất hiện đồng thời của nhiều nội dung tương tự trong khoảng thời gian ngắn hoặc các mẫu lan truyền không phù hợp với hành vi người dùng thông thường có thể cho thấy quá trình tạo sinh tự động hoặc phát tán có tổ chức.

Phân tích dấu vết thiết bị và nguồn dữ liệu

Dữ liệu hình ảnh và video kỹ thuật số thường chứa thông tin về thiết bị thu nhận và quy trình xử lý. Phân tích các dấu vết này suy luận được nguồn gốc thiết bị, phát hiện sự không nhất quán của chuỗi tạo dữ liệu, các dấu hiệu tổng hợp hoặc chỉnh sửa.

Phân tích dấu vết xử lý hậu kỳ

Các thao tác chỉnh sửa và nén dữ liệu có thể để lại các bất nhất cấu trúc trong dữ liệu. Việc truy vết lịch sử xử lý hoặc phát hiện sự không đồng nhất trong đặc trưng nén cung cấp cơ sở bổ sung để nhận diện nội dung bị thao túng.

Phát hiện dựa trên siêu dữ liệu phụ thuộc vào tính toàn vẹn và khả năng truy xuất thông tin phụ trợ. Phương pháp này thường được sử dụng kết hợp với phân tích nội dung và phân tích lan truyền thông tin nhằm nâng cao độ chính xác.

4.3. Phát hiện dựa trên ngữ cảnh mạng xã hội

Trong môi trường mạng xã hội, thông tin sai lệch không tồn tại một cách độc lập mà gắn liền với quá trình lan truyền trong mạng người dùng. Do đó, tin giả có thể được phát hiện thông qua phân tích quá trình lan truyền, cấu trúc mạng lan truyền và hành vi của các tác nhân tham gia [1][6].

Phân tích động học lan truyền

Các nghiên cứu thực nghiệm cho thấy thông tin sai lệch thường có mô hình lan truyền khác biệt so với thông tin đã được kiểm chứng [7]. Việc mô hình hóa quá trình lan truyền cho phép trích xuất các đặc trưng thống kê phục vụ phân loại nội dung.

Phân tích cấu trúc mạng

Mạng xã hội có thể được biểu diễn dưới dạng đồ thị trong đó nút đại diện cho người dùng và cạnh biểu thị quan hệ tương tác. Các phương pháp học trên đồ thị cho phép khai thác cấu trúc lan truyền, nhận diện các cụm cộng đồng có tính khép kín và phát hiện các kết nối bất thường liên quan đến hoạt động phát tán thông tin sai lệch.

Phân tích hành vi tác nhân và phát

hiện tự động hóa

Các tài khoản tự động đóng vai trò quan trọng trong việc khuếch đại phạm vi tiếp cận của nội dung sai lệch. Việc phân tích đặc trưng hành vi, tần suất hoạt động và vị trí cấu trúc của tài khoản trong mạng lưới cho phép nhận diện các tác nhân có dấu hiệu tự động hóa hoặc phối hợp có tổ chức.

Cách tiếp cận dựa trên ngữ cảnh mạng xã hội phụ thuộc vào khả năng truy cập dữ liệu của nền tảng mạng xã hội và thường khó áp dụng trong giai đoạn đầu của quá trình lan truyền khi dữ liệu còn hạn chế.

4.4. Watermark AI và truy vết nguồn gốc

Ngoài các phương pháp phát hiện như nêu trên, một hướng tiếp cận mới là nhúng tín hiệu nhận diện trực tiếp vào nội dung trong quá trình tạo sinh, được gọi là watermark AI. Phương pháp watermark chủ yếu hướng tới truy xuất nguồn gốc tạo sinh của dữ liệu, giúp nhận diện nội dung có được tạo ra bởi hệ thống AI hay không.

Về mặt kỹ thuật, watermark có thể được tích hợp vào quá trình sinh dữ liệu thông qua các cơ chế điều khiển thống kê của mô hình. Với văn bản, watermark thường được thực hiện bằng cách điều chỉnh có kiểm soát phân bố xác suất của các token trong quá trình sinh, từ đó tạo ra các dấu vết thống kê đặc trưng có thể được phát hiện bằng các phép kiểm định phù hợp [13]. Đối với dữ liệu hình ảnh hoặc video, watermark có thể được nhúng dưới dạng tín hiệu ẩn trong miền tần số hoặc trong không gian đặc trưng của mô hình. Tuy nhiên, hiệu quả của watermark phụ thuộc đáng kể vào độ bền của tín hiệu trước các thao tác hậu xử lý như nén dữ liệu, chỉnh sửa nội dung hoặc chuyển đổi định dạng. Những biến đổi này có thể làm suy giảm hoặc loại bỏ tín hiệu watermark, gây khó khăn cho quá trình truy vết nguồn gốc. Ngoài ra, trong môi trường mã nguồn mở hoặc các hệ thống phân tán, việc áp dụng watermark một cách đồng bộ và nhất quán gặp nhiều thách thức.

4.5. So sánh các phương pháp phát hiện tin giả

Các phương pháp phát hiện tin giả hiện nay có những đặc điểm bổ trợ lẫn nhau, không có phương pháp nào đơn lẻ có thể giải quyết toàn diện bài toán.

Phương pháp dựa trên nội dung có ưu điểm là khai thác trực tiếp dữ liệu đầu vào, dễ triển khai và có thể hoạt động ngay từ giai đoạn đầu khi thông tin mới xuất hiện [12]. Phương pháp này có hạn chế khi các mô hình tạo sinh ngày càng tinh vi, làm giảm sự khác biệt giữa nội dung thật và giả.

Phương pháp dựa trên siêu dữ liệu cho phép khai thác các dấu vết ngoại sinh. Ưu điểm của cách tiếp cận này là khả năng phát hiện các bất thường trong quy trình tạo và phân phối nội dung. Nhược điểm lớn là phụ thuộc vào tính toàn vẹn của siêu dữ liệu, trong khi các thông tin này có thể bị xóa hoặc giả mạo.

Phương pháp dựa trên ngữ cảnh mạng xã hội có khả năng phân tích sâu quá trình lan truyền, cấu trúc mạng và hành vi người dùng [1], [12]. Tuy nhiên, nó phụ thuộc vào khả năng truy cập dữ liệu nền tảng và thường không hiệu quả trong giai đoạn sớm khi dữ liệu lan truyền còn hạn chế.

Phương pháp watermark và truy vết có ưu điểm là khả năng xác minh nội dung do AI tạo ra một cách có hệ thống [13]. Tuy nhiên, phương pháp này phụ thuộc vào việc tích hợp từ phía nhà phát triển và gặp thách thức về độ bền tín hiệu trước các biến đổi hậu xử lý.

Bảng 1. So sánh ưu, nhược điểm các phương pháp phát hiện tin giả

Phương pháp	Ưu điểm	Nhược điểm	Phù hợp giai đoạn
Dựa trên nội dung (Content-based)	- Không phụ thuộc nền tảng - Áp dụng sớm khi tin vừa xuất hiện - Hiệu quả với dữ liệu lớn	- Dễ bị đánh lừa bởi nội dung do AI tinh vi - Khó phát hiện khi nội dung "giống thật"	Giai đoạn sớm
Dựa trên siêu dữ liệu (Metadata-based)	- Phát hiện bất thường nguồn gốc - Ít phụ thuộc nội dung	- Dễ bị xóa/sửa metadata - Phụ thuộc khả năng truy cập dữ liệu	Giai đoạn trung gian
Dựa trên ngữ cảnh mạng xã hội (Social-context)	- Phát hiện chiến dịch có tổ chức - Nhận diện bot hiệu quả	- Cần dữ liệu lớn từ nền tảng - Không hiệu quả khi tin chưa lan rộng	Giai đoạn lan truyền
Watermark	- Xác định nội dung do AI tạo - Không phụ thuộc phân tích nội dung	- Phụ thuộc nhà phát triển - Dễ bị suy giảm qua chỉnh sửa - Không đánh giá đúng/sai nội dung	Hậu kiểm/ truy vết

5. THÁCH THỨC VÀ ĐỊNH HƯỚNG NGHIÊN CỨU TRONG PHÁT HIỆN TIN GIẢ

Sự phát triển của GenAI làm gia tăng độ phức tạp của thông tin sai lệch và cung cấp các phương tiện kỹ thuật để phát hiện, phân tích và kiểm chứng nội dung. Tính chất hai mặt này phản ánh đặc trưng song hành của AI trong hệ sinh thái truyền thông số: AI vừa là nguồn phát sinh rủi ro thông tin vừa là thành phần cốt lõi của hạ tầng kiểm chứng tự động.

5.1. Vai trò hỗ trợ kiểm chứng của AI

Phân tích nội dung quy mô lớn: Các hệ thống học sâu cho phép xử lý và phân tích dữ liệu đa phương thức ở quy mô lớn, hỗ trợ nhận diện bất thường trong cấu trúc ngôn ngữ, đặc trưng hình ảnh và tín hiệu âm thanh. Mô hình đa phương thức đặc biệt hữu ích trong phát hiện nội dung tích hợp nhiều dạng biểu đạt.

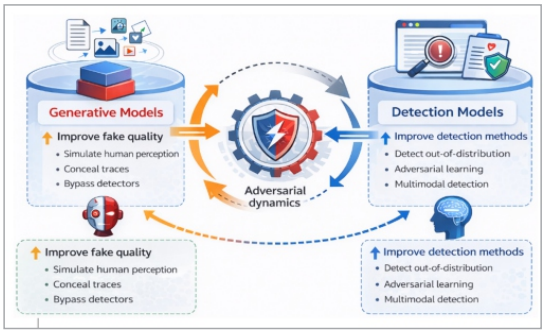
Hỗ trợ quy trình kiểm chứng: AI có thể tự động trích xuất thông tin cần kiểm chứng, truy xuất nguồn tham chiếu và tổng hợp bằng chứng liên quan. Các kiến trúc kết hợp truy xuất và tạo sinh giúp nâng cao hiệu quả xử lý thông tin trong các quy trình kiểm chứng bán tự động.

Phân tích lan truyền: Phương pháp học máy và học trên đồ thị cho phép nhận diện các mẫu lan truyền bất thường, cụm tài khoản phối hợp và hoạt động khuếch đại có chủ đích. AI có thể hỗ trợ phát hiện sớm các chiến dịch thông tin sai lệch ở cấp độ hệ thống.

5.2. Cuộc đua "AI chống AI"

Mối quan hệ giữa mô hình tạo sinh và mô hình phát hiện có thể được xem như một hệ thống đối kháng động. Sự tương tác giữa mô hình tạo sinh và mô hình phát hiện tạo

thành một vòng lặp cạnh tranh về công nghệ. Nó phản ánh bản chất "AI chống AI" trong bài toán kiểm chứng thông tin. Mỗi cải tiến trong năng lực tạo sinh tạo ra áp lực cập nhật đối với hệ thống phát hiện, dẫn đến chu trình tái huấn luyện liên tục. Một thách thức lớn là khả năng tổng quát hóa của mô hình phát hiện đối mặt với dữ liệu nằm ngoài phân phối huấn luyện hoặc các kiến trúc GenAI mới ra đời. Do đó, nhu cầu phát triển các phương pháp phát hiện tin giả dựa trên nguyên lý nền tảng thay vì phụ thuộc vào đặc trưng riêng của từng mô hình.



Hình 2. Động học đối kháng giữa mô hình tạo sinh và phát hiện

Trong Hình 2 minh họa vòng lặp cạnh tranh giữa mô hình tạo sinh (Generative Models) và mô hình phát hiện (Detection Models). Trong đó các hệ thống tạo sinh liên tục cải thiện chất lượng nội dung giả, trong khi các hệ thống phát hiện nâng cao năng lực nhận diện thông qua các kỹ thuật.

5.3. Các thách thức nghiên cứu trong kỷ nguyên GenAI

Bên cạnh những tiến bộ, phát hiện tin giả vẫn đối mặt với nhiều thách thức quan trọng:

Sự tiến hóa nhanh của nội dung giả: Nội dung do AI tạo ra ngày càng tinh vi và thay đổi liên tục khiến các mô hình phát hiện khó thích ứng nếu chỉ dựa trên dữ liệu huấn luyện tĩnh [4]. Tính đa phương thức cũng là một trở ngại lớn trong phát hiện tin giả.

Thiếu dữ liệu chuẩn hóa và gán nhãn: Việc xây dựng bộ dữ liệu đa dạng, đáng tin cậy vẫn là một thách thức lớn trong các nghiên cứu.

Tính giải thích của mô hình (Explainability) [14]: Các mô hình học sâu có độ chính xác cao nhưng khó giải thích, gây hạn chế trong ứng dụng thực tiễn.

Tính hai mặt của AI: Cùng một công nghệ có thể vừa dùng để phát hiện vừa dùng để tạo ra tin giả, làm gia tăng độ phức tạp của bài toán.

5.4. Định hướng nghiên cứu trong tương lai

Từ các phân tích ở trên, một số hướng nghiên cứu tiềm năng có thể được đề xuất cho nghiên cứu phát hiện tin giả trong kỷ nguyên của sự phát triển AI như sau:

- Phát triển mô hình đa phương thức (Multimodal Learning): Kết hợp đồng thời văn bản, hình ảnh, video và âm thanh để nâng cao độ chính xác phát hiện.
- Mô hình lai (Hybrid models): Kết hợp giữa phương pháp truyền thống (rule-based, statistical) và phương pháp học sâu nhằm tận dụng ưu điểm của cả hai.
- Tích hợp LLM với hệ thống kiểm chứng (LLM + RAG): Sử dụng truy xuất tri thức ngoài để kiểm chứng thông tin theo thời gian thực.
- Phát hiện sớm (Early detection): Tập trung vào nhận diện tin giả ngay khi mới xuất hiện, trước khi lan truyền rộng.
- Chuẩn hóa watermark và truy vết nguồn gốc: Xây dựng các tiêu chuẩn kỹ thuật chung nhằm tăng khả năng tương thích giữa các hệ thống.
- Học liên tục: Cho phép mô hình

thích ứng với các dạng tin giả mới mà không cần huấn luyện lại toàn bộ.

6. KẾT LUẬN

GenAI đang làm biến đổi sâu sắc bản chất của tin giả trong môi trường truyền thông số, từ hình thức biểu đạt đến quy mô và cơ chế lan truyền. Nghiên cứu này cho thấy việc phát hiện nội dung sai lệch không thể dựa vào một nguồn tín hiệu đơn lẻ mà cần tiếp cận tích hợp, kết hợp giữa phân tích nội dung, siêu dữ liệu, ngữ cảnh lan truyền và truy vết nguồn gốc. Đồng thời, mối quan hệ đối kháng giữa mô hình tạo sinh và mô hình phát hiện cho thấy bài toán kiểm chứng thông tin là một bài toán động. Nó đòi hỏi phương pháp có khả năng tổng quát hóa cao. Do đó, cần phát triển các khung phát hiện đa tầng và có cơ chế quản trị công nghệ phù hợp nhằm duy trì tính toàn vẹn của hệ sinh thái thông tin trong kỷ nguyên AI.

TÀI LIỆU THAM KHẢO

[1] Truong Hai Bang, Tran Van Cuong (2023), A framework for fake news detection based on the wisdom of crowds and the ensemble learning model, *Computer Science and Information Systems*, 20(4), 1439-1457.

[2] Bommasani R. et al. (2021), On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258.

[3] Dhariwal, P., & Nichol, A.(2021), Diffusion models beat gans on image synthesis, Advances in neural information processing systems, 34, 8780-8794.

[4] Feuerriegel, S. et al. (2024), Generative AI. Business & Information Systems Engineering, 66, 111-126, doi.org/10.1007/s12599-023-00834-7

[5] Li X. et al. (2025), A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective, *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(4), 2658-2675, doi: 10.1109/TETCI.2025.3543389.

[6] Shu, K. et al.(2020), Combating disinformation in a social media age. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6).

[7] Zhou, X., & Zafarani, R. (2020), A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR), 53(5), 1-40.

[8] Tolosana, R. Et al. (2020), Deepfakes and beyond: A survey of face manipulation and fake detection. *Information fusion*, 64, 131-148.

[9] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023), Detectgpt: Zero-shot machine-generated text detection using probability curvature. In International conference on machine learning (pp. 24950-24962), PMLR.

[10] Shaib, C., Elazar, Y., Li, J. J., & Wallace, B. C.(2024), Detection and measurement of syntactic templates in generated text. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 6416-6431).

[11] O'Sullivan, J. (2025), Stylometric comparisons of human versus AI-generated creative writing, *Humanities and Social Sciences Communications*, 12(1), 1-6.

[12] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017), Fake news detection on

social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1), 22-36.

[13] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T.(2023), A watermark for large language models, In International conference on machine learning (pp.17061-17084), PMLR.

[14] Kondamudi, M., & ranjan Sahoo, S.(2025), Integrating explainable AI with enhanced ensemble models for accurate and transparent fake news detection in osn's, *Procedia Computer Science*, 258, 1081-1090.

Lời cảm ơn: *Nghiên cứu này được tài trợ bởi đề tài khoa học và công nghệ cấp cơ sở của Trường Đại học Quảng Bình, mã số đề tài: CS.01.2026*

Liên hệ:
TS. Trần Văn Cường
Khoa Công nghệ thông tin, Trường Đại học Quảng Bình
Địa chỉ: 18 Nguyễn Văn Linh, phường Đồng Thuận, tỉnh Quảng Trị
Email: vancuongqbuni@gmail.com

Ngày nhận bài: 26/02/2026
Ngày gửi phản biện: 26/02/2026
Ngày duyệt đăng: 15/3/2026