Tạp chí Lý Luận Chính Trị Và Truyên Thông

Tạp chí điện tử tiếng Việt - Tập 32, Số 9 (2026)

Ứng dụng trí tuệ nhân tạo (AI) trong kiểm soát tin giả trên TikTok

Ngày nhận bài: 12/03/2026 | **Ngày phản biện:** 21/05/2026 | **Ngày xuất bản:** 10/09/2026

Tác giả: Nguyễn Văn Thiệu (Khoa Truyền thông đa phương tiện, trường Đại học Thăng Long,); Hoàng Thị Từ Quy (Học viên cao học Quản lý Báo chí Truyền thông K30.2A, Học viện Báo chí và Tuyên truyền)

Tóm tắt: Bài viết phân tích vai trò của trí tuệ nhân tạo trong kiểm soát tin giả trên nền tảng TikTok trong bối cảnh thông tin sai lệch ngày càng lan truyền mạnh trên các nền tảng mạng xã hội. Tác giả sử dụng phương pháp nghiên cứu tài liệu thông qua việc tổng hợp, hệ thống hóa và phân tích các công trình nghiên cứu trong và ngoài nước, báo cáo của TikTok cùng các tài liệu liên quan đến ứng dụng AI trong phát hiện và kiểm soát tin giả. Kết quả nghiên cứu cho thấy AI đang được TikTok ứng dụng thông qua các công nghệ như xử lý ngôn ngữ tự nhiên, học máy, học sâu nhằm phát hiện, sàng lọc và hạn chế sự lan truyền của thông tin sai lệch. Tuy nhiên, hiệu quả của các hệ thống này vẫn chịu tác động bởi sự phát triển của công nghệ deepfake, hạn chế của dữ liệu tiếng Việt, đặc điểm vận hành của thuật toán đề xuất và thói quen tiếp nhận thông tin của người dùng. Trên cơ sở đó, bài viết đề xuất một số giải pháp nhằm nâng cao hiệu quả kiểm soát tin giả trên TikTok như hoàn thiện công nghệ AI, tăng cường phối hợp giữa các chủ thể liên quan và nâng cao năng lực truyền thông số của người dùng.

1. Đặt vấn đề

Sự phát triển của các nền tảng truyền thông xã hội đã làm thay đổi căn bản cách thức sản xuất, phân phối và tiếp nhận thông tin. Trong đó, TikTok với đặc trưng video ngắn và hệ thống đề xuất nội dung cá nhân hóa dần trở thành một kênh quan trọng để người dùng tiếp cận thông tin. "Tính đến năm 2025, TikTok có 40,9 triệu người dùng từ 18 tuổi trở lên tại Việt Nam" (DataReportal, 2025). Tuy nhiên, "khả năng lan truyền mạnh mẽ và sức hấp dẫn của nội dung trên TikTok biến nền tảng này trở thành môi

trường thuận lợi cho việc phát tán nhanh chóng tin giả" (Lan & Tung, 2024, tr.3).

Trong bối cảnh đó, trí tuệ nhân tạo bắt đầu được sử dụng như một công cụ hỗ trợ phát hiện và kiểm soát các thông tin sai lệch, tin giả. "Việc phát hiện tin giả trên mạng xã hội không thể chỉ dựa vào nội dung của bản tin mà cần xem xét cả **nội dung thông tin và bối cảnh xã hội"** (Zhou & Zafarani, 2020, tr.31), bởi các đặc điểm về tương tác xã hội và hành vi của người dùng cũng cung cấp những thông quan trọng cho quá trình nhận diện tin giả. Đối với môi trường mạng xã hội, tin giả không chỉ tồn tại dưới dạng văn bản mà còn có thể được truyền tải qua hình ảnh, video và âm thanh. Do đó, "trí tuệ nhân tạo (AI), mô hình ngôn ngữ lớn (LLM) và phân tích mạng xã hội cũng đóng vai trò quan trọng trong phát hiện tin giả đa phương thức" (Lv et al., 2025, tr.1).

Tuy nhiên, hiệu quả của việc ứng dụng AI trong phát hiện tin giả vẫn còn nhiều giới hạn. "Các sự kiện mới nổi thường tạo ra thông tin và kiến thức mới nên hệ thống gặp khó khăn khi thông tin chưa được cập nhật kịp thời" (Zhou & Zafarani, 2020, tr31.). Ngoài ra, "người dùng còn phải đối mặt với nguy cơ từ tin giả và thông tin sai lệch do tính năng đề xuất nội dung tự động" (Sharma et al., 2024, tr1.). Như vậy, kiểm soát tin giả không chỉ dừng lại ở việc phát hiện nội dung sai lệch mà còn phải xem xét khả năng nội dung đó có thể tiếp tục được thuật toán phân phối đến người dùng.

Vì vậy, nghiên cứu này là cần thiết nhằm bổ sung cơ sở lý luận và đưa ra giải pháp nâng cao hiệu quả trong việc kiểm soát tin giả trong môi trường số hiện nay. Nghiên cứu tập trung trả lời ba câu hỏi: (1) AI được ứng dụng như thế nào trong phát hiện và kiểm soát tin giả trên TikTok? (2) Những yếu tố nào hạn chế hiệu quả của AI trong quá trình này? (3) Cần có những giải pháp nào để nâng cao hiệu quả ứng dụng AI trong kiểm soát tin giả trên TikTok?

2. Phương pháp nghiên cứu

Bài viết sử dụng phương pháp nghiên cứu tài liệu, kết hợp với phân tích, tổng hợp và so sánh các nguồn tư liệu nhằm làm rõ việc ứng dụng trí tuệ nhân tạo trong kiểm soát tin giả trên TikTok. Nguồn tư liệu được lựa chọn gồm ba nhóm chính. Thứ nhất là các công trình nghiên cứu học thuật về tin giả, AI, phát hiện thông tin sai lệch và kiểm chứng thông tin. Thứ hai là các tài liệu của TikTok liên quan đến chính sách thông tin sai lệch, cơ chế kiểm duyệt, hệ thống đề xuất và việc ứng dụng công nghệ trong kiểm

soát nội dung. Thứ ba là các tài liệu của cơ quan quản lý và cơ quan báo chí Việt Nam, được sử dụng để xem xét những trường hợp cụ thể.

Quá trình nghiên cứu được thực hiện theo ba bước. Bước thứ nhất, thu thập và hệ thống hóa các tài liệu theo các từ khóa và chủ đề liên quan. Bước thứ hai, phân tích và đối chiếu nội dung các tài liệu theo ba nhóm vấn đề: (1) cơ chế và công nghệ AI được sử dụng trong phát hiện, sàng lọc và xử lý thông tin sai lệch; (2) những thách thức đối với hiệu quả kiểm soát tin giả trên TikTok; (3) cơ chế phối hợp giữa nền tảng, con người, tổ chức kiểm chứng và cơ quan quản lý. Bước thứ ba, tổng hợp các kết quả và so sánh giữa các nguồn để xác định những vấn đề còn hạn chế, từ đó đề xuất các giải pháp phù hợp với bối cảnh Việt Nam.
Phạm vi nghiên cứu tập trung vào ứng dụng AI trong kiểm soát tin giả trên nền tảng TikTok, không nhằm đánh giá độ chính xác kỹ thuật của một mô hình AI cụ thể và cũng không tiến hành thu thập, mã hóa hay định lượng mẫu video TikTok. Các trường hợp thực tế được sử dụng trong bài có chức năng minh họa và đối chiếu cho những vấn đề được xác định từ hệ thống tài liệu. Vì vậy, kết quả nghiên cứu được hiểu theo hướng phân tích và tổng hợp lý luận – thực tiễn, không phải kết quả của một nghiên cứu thực nghiệm trên dữ liệu video TikTok.

3. Kết quả và Thảo luận

3.1. Kết quả nghiên cứu

3.1.1. Thực trạng ứng dụng trí tuệ nhân tạo trong kiểm soát tin giả trên TikTok

Cơ chế ứng dụng trí tuệ nhân tạo trong kiểm soát tin giả trên TikTok

Việc phát hiện tin giả không chỉ dựa vào việc phân tích nội dung văn bản mà còn cần xem xét các đặc điểm của nguồn tin, hành vi người dùng và quá trình lan truyền. Nghĩa là, "đặc điểm của thông tin và bối cảnh xã hội đều có ý nghĩa đối với việc xác định tin giả" (Shu et al., 2017, tr.23). "Các phương pháp phát hiện tin giả cũng được phát triển theo nhiều hướng, từ phân tích nội dung, phong cách ngôn ngữ đến đặc điểm của nguồn tin và mạng lưới lan truyền" (Zhou & Zafarani, 2020, tr.4). Cách tiếp cận này tạo cơ sở để AI có thể xử lý đồng thời nhiều loại tín hiệu nhằm hỗ trợ phát hiện nội dung có dấu hiệu sai lệch.

Trong thực tiễn, các công nghệ học máy và học sâu được sử dụng để tự động hóa quá trình sàng lọc nội dung. Theo TikTok, "nền tảng sử dụng các mô hình học máy để hỗ trợ phát hiện thông tin sai lệch tiềm ẩn" (Keenan, 2022). Việc sử dụng mô hình tự động có vai trò quan trọng khi TikTok phải xử lý một lượng lớn video được tải lên liên tục. Thay vì chờ người dùng báo cáo từng nội dung, hệ thống tự động có thể xác định trước những nội dung có dấu hiệu cần xem xét, từ đó chuyển sang các bước kiểm tra tiếp theo. TikTok cũng biết "nền tảng tiếp tục đầu tư vào các mô hình học máy và làm sao để khả năng cập nhật mô hình này nhanh hơn bản chất thay đổi nhanh chóng của thông tin sai lệch" (Keenan, 2022).

Khác với các nền tảng truyền thông chủ yếu dựa trên văn bản, TikTok truyền tải thông tin thông qua sự kết hợp của hình ảnh, video, âm thanh, lời nói, chữ viết trên màn hình. Vì vậy, việc phát hiện thông tin sai lệch đòi hỏi hệ thống không chỉ phân tích ngôn ngữ mà còn phải nhận diện các dấu hiệu bất thường trong hình ảnh và âm thanh. "Nền tảng đầu tư vào việc cải thiện khả năng phát hiện âm thanh và hình ảnh gây hiểu lầm để giảm nội dung sai lệch" (Keenan, 2022). Như vậy, AI được mở rộng từ việc phân tích nội dung văn bản sang nhận diện những hình ảnh hoặc âm thanh có khả năng đã bị thao túng.

TikTok phát triển các công nghệ nhằm nhận diện nội dung AI và bắt đầu sử dụng Chứng chỉ Nội dung (Content Credentials) để hỗ trợ xác định nguồn gốc của nội dung. Trong báo cáo năm 2025, TikTok cho biết nền tảng đã triển khai khả năng đọc Chứng chỉ Nội dung để "tự động gắn nhãn nội dung do AI tạo có nguồn gốc trên các nền tảng chính khác" (TikTok, 2025). Cơ chế này cho phép nền tảng bổ sung thông tin về nguồn gốc nội dung, từ đó giúp người dùng nhận biết những nội dung đã được tạo hoặc chỉnh sửa bằng AI.

Sự phát triển gần đây của các mô hình ngôn ngữ lớn cũng đang mở rộng khả năng ứng dụng AI trong kiểm soát tin giả trên TikTok. "TikTok sử dụng Mô hình ngôn ngữ lớn (LLMs) để hỗ trợ kiểm duyệt chủ động ở quy mô lớn. Theo đó, LLMs có thể hiểu ngôn ngữ con người và thực hiện các nhiệm vụ phức tạp" (TikTok, 2025).

Tuy nhiên, cũng cần phân biệt giữa khả năng phát hiện của AI và khả năng xác định tính đúng – sai của thông tin. Chính TikTok thừa nhận rằng thông tin sai lệch khác với

nhiều loại nội dung vi phạm khác bởi "ngữ cảnh và việc xác minh tính xác thực là rất quan trọng để thực thi một cách nhất quán và chính xác các chính sách về thông tin sai lệch" (Keenan, 2022).

Mô hình kiểm soát tin giả nhiều lớp trên TikTok

Kết quả nghiên cứu cho thấy việc kiểm soát tin giả trên TikTok được thực hiện không chỉ bằng một công nghệ hoặc một chủ thể riêng lẻ mà dựa trên sự kết hợp giữa hệ thống phát hiện tự động, kiểm duyệt của con người, kiểm chứng thông tin và các biện pháp can thiệp của nền tảng. Một nội dung có thể chứa thông tin đúng nhưng được đặt trong một bối cảnh sai, hoặc sử dụng hình ảnh và phát biểu có thật để tạo ra một diễn giải sai lệch. Vì vậy, khả năng nhận diện mẫu của AI cần được kết hợp với khả năng đánh giá ngữ cảnh và xác minh nguồn tin.

TikTok cũng xác định rõ giới hạn của hệ thống tự động trong quá trình xử lý thông tin sai lệch. "Chúng tôi sử dụng các mô hình học máy để giúp phát hiện thông tin sai lệch tiềm ẩn", nhưng "cuối cùng thì cách tiếp cận của chúng tôi ngày nay là yêu cầu nhóm kiểm duyệt đánh giá, xác nhận và xóa các vi phạm thông tin sai lệch" (Keenan, 2022). Như vậy, AI trước hết thực hiện chức năng phát hiện những nội dung có khả năng chứa thông tin sai lệch, sau đó đội ngũ kiểm duyệt tiếp tục đánh giá, xác nhận và đưa ra quyết định xử lý. Cơ chế này cho thấy AI không thay thế hoàn toàn con người mà đóng vai trò như một lớp sàng lọc ban đầu nhằm nâng cao tốc độ và quy mô xử lý nội dung.

Vai trò kiểm tra trình duyệt của con người đặc biệt quan trọng đối với những nội dung cần xem xét về bối cảnh. "Ngữ cảnh và việc xác minh tính xác thực là rất quan trọng để thực thi một cách nhất quán và chính xác các chính sách về thông tin sai lệch" (Keenan, 2022). Nghĩa là, một hệ thống AI có thể nhận diện từ khóa, hình ảnh hoặc những đặc điểm bất thường nhưng chưa đủ để đưa ra kết luận cuối cùng về tính xác thực của một tuyên bố. Người kiểm duyệt cần xem xét nội dung trong mối quan hệ với thời điểm, sự kiện, nhân vật, nguồn phát và các thông tin liên quan trước khi đưa ra quyết định.

Việc kiểm soát tin giả trên TikTok không chỉ dựa vào hệ thống AI mà còn được thực hiện thông qua cơ chế phối hợp giữa nền tảng và cơ quan quản lý nhà nước. "Năm

2024, Cục Phát thanh, truyền hình và thông tin điện tử (Bộ TT&TT) tiếp tục duy trì việc kết hợp các giải pháp đấu tranh cứng rắn, linh hoạt với các nền tảng xuyên biên giới, yêu cầu các nền tảng phải ngăn chặn, gỡ bỏ thông tin vi phạm pháp luật Việt Nam" (Cục Phát thanh, truyền hình và thông tin điện tử, 2025). Theo đó, "tính đến tháng 12/2024, TikTok đã chặn, gỡ 971 nội dung vi phạm, bao gồm 677 video và 294 tài khoản" (Cục Phát thanh, truyền hình và thông tin điện tử, 2025).

3.1.2. Thách thức trong việc kiểm soát tin giả trên TikTok

Dù trí tuệ nhân tạo giữ vai trò quan trọng trong hệ thống kiểm soát nội dung của TikTok, song vẫn tồn tại nhiều thách thức khiến tin giả khó bị loại bỏ hoàn toàn.

Thứ nhất, thách thức về mặt công nghệ: "TikTok đã trở thành nền tảng nơi các video lan truyền thường xuất hiện đầu tiên" (Ling et al., 2021, tr.1). Một video chỉ dài vài chục giây nhưng có thể đạt hàng nghìn lượt xem trong thời gian rất ngắn nếu trùng với xu hướng hoặc được thuật toán đưa vào luồng đề xuất. Bên cạnh đó, công nghệ deepfake, voice clone và các kỹ thuật cắt ghép tinh vi có thể tạo ra những video với độ chân thực cao, làm giảm hiệu quả của các mô hình phát hiện tự động. Khi độ chính xác của deepfake cao, các phương pháp dựa trên nhận diện khuôn mặt hoặc giọng nói trở nên kém hiệu quả, khiến nhiều video sai sự thật chỉ được phát hiện sau khi đã lan truyền rộng rãi. Một người dùng TikTok đã cắt ghép chèn nội dung bịa đặt về ông Dương Khắc Mai đang phát biểu tại Quốc hội với tiêu đề "Đề xuất sốc: vợ hoặc chồng đi nhậu quá 20 giờ có thể bị xử phạt nghiêm, mức phạt dự kiến lên đến 30 triệu đồng" (Báo Thanh Niên, 2025). Video đã thu hút đến một triệu lượt xem trên TikTok và hàng chục nghìn lượt tương tác. Trường hợp này cho thấy, AI vẫn chưa thể xác minh chính xác một số nội dung được tạo ra từ việc cắt ghép, thay đổi ngữ cảnh dữ liệu do nội dung vẫn có hình ảnh thật, nhân vật thật.

Thứ hai, thách thức về ngôn ngữ và bộ dữ liệu tiếng Việt: Tiếng Việt là ngôn ngữ giàu sắc thái, đa nghĩa và ngày càng phong phú hơn khi mạng xã hội phát triển. Nhiều nội dung trên TikTok còn sử dụng từ viết tắt, tiếng lóng, meme, ký hiệu hoặc các cách diễn đạt mang tính hàm ý, khiến ý nghĩa của thông tin phụ thuộc nhiều vào ngữ cảnh và đặc điểm văn hóa của cộng đồng người dùng. Những hạn chế này khiến AI khó giải mã ngữ nghĩa chính xác, dẫn đến hiểu sai bối cảnh hoặc không nhận diện được các

dạng nội dung "nửa thật nửa giả". "Phát hiện tin giả gặp nhiều khó khăn vì thông tin về ngữ cảnh xã hội chưa có hoặc chưa đủ" (Zhou & Zafarani, 2020, tr.32). Ngoài ra, nguồn dữ liệu tiếng Việt phục vụ huấn luyện các mô hình AI còn chưa thực sự phong phú và được cập nhật thường xuyên so với các ngôn ngữ phổ biến. Do đó, khả năng học hỏi và thích ứng của mô hình AI bị giảm, không thích ứng kịp với các xu hướng ngôn ngữ mới trên mạng xã hội và ảnh hưởng đến độ chính xác trong việc phát hiện tin giả.

Thứ ba, thách thức về thuật toán đề xuất cá nhân hoá của TikTok: TikTok sử dụng thuật toán đề xuất cá nhân hóa nhằm tối ưu hóa trải nghiệm người dùng thông qua việc phân tích lịch sử xem, lượt thích, bình luận, chia sẻ và các hành vi tương tác khác. Người dùng có thể nhanh chóng tiếp cận các nội dung phù hợp với sở thích cá nhân nhưng cũng có nguy cơ tương tác với các thông tin chưa được kiểm chứng. Bởi vì "quá trình lan truyền video chịu ảnh hưởng bởi đặc điểm nội dung cũng như hành vi tương tác của người dùng" (Ling et al., 2021, tr.1). Trong các tình huống nhạy cảm như thiên tai, dịch bệnh hoặc các sự kiện chính trị - xã hội, những nội dung gây tò mò, tranh cãi hoặc kích thích cảm xúc thường thu hút lượng tương tác lớn và có xu hướng được thuật toán ưu tiên phân phối. Do đó, thông tin sai lệch có thể lan truyền với tốc độ rất nhanh trước khi được hệ thống AI hoặc đội ngũ kiểm duyệt phát hiện và xử lý.

Thứ tư, thách thức về hành vi tiếp nhận thông tin của người dùng: Hiệu quả kiểm soát tin giả không chỉ phụ thuộc vào năng lực của hệ thống AI mà còn chịu tác động đáng kể từ hành vi tiếp nhận thông tin của người dùng. Các video TikTok thường ngắn, nội dung cô đọng và liên tục đề xuất khiến người dùng có xu hướng tiếp nhận thông tin nhanh, dựa nhiều vào cảm xúc mà không kiểm chứng thông tin. Theo Vázquez-Herrero, "Người dùng tiêu thụ thông tin một cách ngẫu nhiên mà không cần phải chủ động tìm kiếm cũng như không dành nhiều thời gian để phân tích hoặc kiểm chứng" (Vázquez-Herrero et al., 2022, tr.3829). Ngoài ra, "mức độ tin tưởng của người dùng TikTok còn chịu ảnh hưởng bởi uy tín của người sáng tạo nội dung và sự quen thuộc với tài khoản đăng tải" (Lan & Tung, 2024, tr.5). Vì vậy, khi năng lực hiểu biết truyền thông của người dùng còn hạn chế, AI chỉ có thể đóng vai trò là công cụ hỗ trợ mà chưa thể thay thế hoàn toàn khả năng đánh giá và sàng lọc thông tin của con người.

3.2. Thảo luận

Kết quả nghiên cứu cho thấy trí tuệ nhân tạo đang làm thay đổi cách thức kiểm soát tin giả trên TikTok, từ phương thức phụ thuộc chủ yếu vào phát hiện và xử lý thủ công sang cơ chế kết hợp giữa tự động hóa và kiểm duyệt của con người. AI có khả năng xử lý khối lượng lớn dữ liệu, phát hiện các dấu hiệu bất thường và hỗ trợ sàng lọc nội dung trong thời gian ngắn. Kết quả này phù hợp với sự phát triển của các nghiên cứu về phát hiện tin giả, từ những phương pháp dựa trên nội dung và nguồn tin đến các phương pháp khai thác bối cảnh xã hội, mô hình lan truyền và dữ liệu đa phương thức. Đặc biệt, sự xuất hiện của LLM và các mô hình đa phương thức cho thấy AI đang chuyển từ chức năng nhận diện những đặc điểm bề mặt của nội dung sang hỗ trợ phân tích các tuyên bố và mối quan hệ giữa nhiều dạng dữ liệu khác nhau.

Để nâng cao hiệu quả kiểm soát tin giả trên TikTok, cần triển khai đồng bộ các giải pháp về công nghệ, quản lý và giáo dục với sự tham gia của nhiều chủ thể:

Thứ nhất, nền tảng TikTok cần tiếp tục hoàn thiện hệ thống AI theo hướng nâng cao khả năng xử lý tiếng Việt và phát hiện các dạng thông tin sai lệch mới như deepfake, voice clone hoặc nội dung được chỉnh sửa bằng AI tạo sinh. Bên cạnh việc nâng cấp các mô hình xử lý ngôn ngữ tự nhiên và thị giác máy tính, TikTok cần xây dựng bộ dữ liệu tiếng Việt có quy mô lớn và được cập nhật thường xuyên nhằm nâng cao độ chính xác của hệ thống phát hiện tự động. Đồng thời, nền tảng cần tăng cường tính minh bạch trong hoạt động kiểm duyệt thông qua việc hiển thị rõ lý do gắn nhãn cảnh báo, hạn chế phân phối hoặc gỡ bỏ nội dung; hoàn thiện cơ chế kháng nghị đối với người sáng tạo nội dung; đồng thời mở rộng hợp tác với các tổ chức kiểm chứng thông tin độc lập nhằm nâng cao độ tin cậy của quá trình xác minh thông tin.

Thứ hai, các cơ quan quản lý nhà nước cần tiếp tục hoàn thiện cơ chế phối hợp với các nền tảng trong việc phát hiện và xử lý thông tin sai lệch. Cơ chế này cần hướng đến việc chia sẻ thông tin cảnh báo sớm về các nội dung có nguy cơ lan truyền rộng, thiết lập quy trình phối hợp trong xác minh và yêu cầu gỡ bỏ nội dung vi phạm, đồng thời tăng cường trao đổi dữ liệu phục vụ công tác quản lý theo quy định của pháp luật. Đối với các vụ việc có phạm vi ảnh hưởng lớn, cần có sự phối hợp giữa cơ quan quản lý, nền tảng số và các cơ quan báo chí nhằm kịp thời cung cấp thông tin chính xác và giảm nguy cơ lan truyền tin giả trên nhiều nền tảng khác nhau.

Thứ ba, các cơ quan báo chí cần phát huy vai trò là nguồn cung cấp thông tin chính thống, chính xác và đáng tin cậy trong môi trường truyền thông số. Báo chí không chỉ có chức năng phát hiện, phản bác và kiểm chứng các thông tin sai lệch mà còn cần chủ động giải thích những vấn đề phức tạp, phổ biến kiến thức về nhận diện tin giả cho công chúng. Việc tăng cường hợp tác giữa báo chí với TikTok và các tổ chức kiểm chứng thông tin cũng góp phần hỗ trợ hệ thống AI cải thiện khả năng nhận diện và xử lý các nội dung sai lệch.

Thứ tư, các cơ sở giáo dục và người dùng cần coi trọng việc phát triển năng lực truyền thông số. Đây là giải pháp lâu dài nhằm giảm thiểu tác động của tin giả. Các nội dung giáo dục năng lực số, kỹ năng kiểm chứng thông tin có thể được xem xét đưa vào chương trình đào tạo. Đồng thời, cần trang bị cho người học khả năng nhận diện các hình thức thao túng thông tin bằng AI như deepfake, voice clone hoặc các nội dung được tạo sinh bằng trí tuệ nhân tạo. Chỉ khi người dùng có đủ năng lực đánh giá và kiểm chứng thông tin AI mới phát huy hiệu quả trong kiểm soát tin giả trên môi trường số.

4. Kết luận

Nghiên cứu cho thấy trí tuệ nhân tạo đang đóng vai trò ngày càng quan trọng trong việc phát hiện và kiểm soát tin giả trên các nền tảng mạng xã hội. Việc ứng dụng các công nghệ như xử lý ngôn ngữ tự nhiên, học máy, học sâu và thị giác máy tính cho phép nền tảng TikTok phân tích khối lượng dữ liệu lớn, phát hiện sớm các nội dung có dấu hiệu sai lệch và hạn chế tin giả. Tuy nhiên, hiệu quả của các hệ thống phát hiện tin giả dựa trên AI vẫn còn chịu ảnh hưởng bởi nhiều yếu tố như công nghệ deepfake, sự hạn chế của dữ liệu tiếng Việt, xu hướng tiếp nhận thông tin nhanh và thiên về cảm xúc của người dùng. Vì vậy, việc kiểm soát tin giả trên TikTok không thể chỉ dựa vào các giải pháp công nghệ mà cần có sự kết hợp giữa hệ thống phát hiện tự động bằng AI, hoạt động kiểm duyệt của con người, sự hợp tác với các cơ quan, tổ chức khác và đặc biệt là nâng cao năng lực truyền thông số của người dùng. Trong đó, việc phát triển kỹ năng tiếp nhận và đánh giá thông tin được xem là yếu tố quan trọng nhằm hạn chế sự lan truyền và tác động tiêu cực của tin giả./.

Tài liệu tham khảo

Báo Thanh Niên. (2025). Đại biểu Quốc hội Lâm Đồng bị gán ghép hình ảnh với nội dung sai sự thật. https://thanhnien.vn/dai-bieu-quoc-hoi-lam-dong-bi-gan-ghep-hinh-anh-voi-noi-dung-sai-su-that-185250903115041458.htm

Cục Phát thanh, truyền hình và thông tin điện tử. (2025). Gần 16.000 nội dung vi phạm đã được gỡ bỏ trên các nền tảng xuyên biên giới. https://abei.gov.vn/thong-tin-dien-tu/gan-16000-noi-dung-vi-pham-da-duoc-go-bo-tren-cac-nen-tang-xuyen-bien-gioi/118629

DataReportal. (2025). Digital 2025: Vietnam—DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2025-vietnam

Keenan, C. (2022). An update on our work to counter misinformation. Newsroom | TikTok. https://newsroom.tiktok.com/an-update-on-tiktoks-work-to-counter-misinformation?lang=en-GB

Lan, D. H., & Tung, T. M. (2024). Exploring fake news awareness and trust in the age of social media among university student TikTok users. Cogent Social Sciences, 10(1), 2302216. https://doi.org/10.1080/23311886.2024.2302216

Ling, C., Blackburn, J., Cristofaro, E. D., & Stringhini, G. (2021). Slapping Cats, Bopping Heads, and Oreo Shakes: Understanding Indicators of Virality in TikTok Short Videos (arXiv:2111.02452). arXiv. https://doi.org/10.48550/arXiv.2111.02452

Lv, J., Gao, Y., Li, L., Shi, L., & Li, S. (2025). Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. Journal of King Saud University Computer and Information Sciences, 37(9), 306. https://doi.org/10.1007/s44443-025-00317-7

Sharma, A., Pujari, T., & Goel, A. (2024). The Fake Propaganda on Tiktok: Measuring Gen Z's Vulnerability and Building. IJFMR - International Journal For Multidisciplinary Research, 6(5). https://doi.org/10.36948/ijfmr.2024.v06i05.42635

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. https://doi.org/10.1145/3137597.3137600

TikTok. (2025). TikTok Code of Practice on Disinformation: January to June 2025. https://disinfocode.eu/reports/download/155

Vázquez-Herrero, J., Negreira-Rey, M. C., & Sixto-García, J. (2022). (PDF) Young People and Social Networks: News Consumption Habits and Credibility of the News. International Journal of Communication, 16, 3822–3842. https://doi.org/10.58262/V32I78.13

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys (CSUR), 53(5), 109:1-109:40. https://doi.org/10.1145/3395046