RESEARCHING SOLUTIONS AND TOOLS TO DETECTING FAKE NEWS ON SOCIAL MEDIA

NGHIÊN CỨU VỀ CÁC GIẢI PHÁP VÀ CÔNG CỤ PHÁT HIỆN TIN TỨC GIẢ TRÊN MẠNG XÃ HỘI

Hoàng Đình Tuyền[1], Nguyễn Thị Bích Hằng[2], Trần Văn Cường[1]
[1]Khoa Công nghệ thông tin, Trường Đại học Quảng Bình,
[2]Phòng KHCN - ĐN & HL, Trường Đại học Quảng Bình

ABSTRACT: *Over the past decade, social media platforms and blogging services such as Facebook, Twitter, Instagram, and Sina Weibo have evolved into indispensable tools in modern social life. However, the rapid expansion of social networks has also fostered an environment conducive to fraudulent activities, propaganda, and rumor propagation, making users increasingly susceptible to misinformation. Consequently, the detection and prevention of fake news have become an urgent necessity. This paper synthesizes recent studies on fake news detection in social media. We analyze the concept of fake news and related terminologies while introducing publicly available datasets and real-time verification tools. Additionally, the study explores content-based and social context-based approaches for fake news detection, comparing the effectiveness of existing techniques.*

Keywords: *Fake news detection, rumor detection, fact-checking, misinformation, deceptive information.*

TÓM TẮT: *Trong suốt thập kỷ qua, các nền tảng mạng xã hội và các dịch vụ blog như Facebook, Twitter, Instagram và Sina Weibo đã phát triển thành những công cụ không thể thiếu trong đời sống xã hội hiện đại. Tuy nhiên, sự phát triển mạnh mẽ của mạng xã hội cũng tạo ra môi trường thuận lợi cho các hoạt động lừa đảo, tuyên truyền và lan truyền tin đồn, dẫn đến việc người dùng dễ dàng bị đánh lừa. Chính vì thế, việc phát hiện và ngăn chặn tin giả đã trở thành một nhu cầu cấp thiết. Bài báo này tổng hợp các nghiên cứu gần đây liên quan đến các phương pháp phát hiện tin giả trên mạng xã hội. Chúng tôi sẽ phân tích khái niệm tin giả và các thuật ngữ liên quan, đồng thời giới thiệu các bộ dữ liệu công khai và công cụ trực tuyến giúp xác minh tin tức theo thời gian thực. Bài báo cũng trình bày các phương pháp phát hiện tin giả dựa trên nội dung và bối cảnh xã hội, đồng thời so sánh hiệu quả của các kỹ thuật phát hiện tin giả hiện có.*

Từ khoá: *Phát hiện tin giả, phát hiện tin đồn, kiểm tra sự thật, thông tin sai lệch, thông tin đánh lừa.*

1. ĐẶT VẤN ĐỀ

Hiện nay, mạng xã hội đóng vai trò quan trọng trong việc truyền tải thông tin, cho phép người dùng tiếp nhận và chia sẻ nội dung một cách nhanh chóng. Tuy nhiên, sự phát triển mạnh mẽ của các nền tảng này cũng kéo theo sự gia tăng của tin giả, gây ra những tác động tiêu cực đến nhận thức cộng đồng. Do đó, việc nhận diện và kiểm soát tin giả trên mạng xã hội trở thành một nhiệm vụ quan trọng nhằm đảm bảo tính chính xác của thông tin.

Người dùng mạng xã hội thường thiếu kinh nghiệm trong việc đánh giá tính xác thực của thông tin, khiến họ dễ bị tác động và vô tình lan truyền nội dung sai lệch. Một

số nhà phân tích chính trị cho rằng kết quả cuộc bầu cử Tổng thống Mỹ năm 2016, với chiến thắng của Donald Trump, có thể đã bị ảnh hưởng bởi các chiến dịch tuyên truyền và lan truyền tin đồn trên mạng xã hội [1].

Gần đây, sự bùng phát của đại dịch COVID-19 đã làm gia tăng sự lan truyền tin giả trên các nền tảng trực tuyến, gây ra những hậu quả nghiêm trọng [2]. Chúng ta đang sống trong thời đại số, nơi thông tin không chỉ được tiếp nhận mà còn được tạo ra bởi chính người dùng. Tuy nhiên, phần lớn thông tin không được kiểm chứng do thiếu các công cụ xác minh đáng tin cậy. Vì vậy, nghiên cứu về phát hiện tin giả trên mạng xã hội đang thu hút sự quan tâm lớn từ cộng đồng học thuật cũng như các công ty công nghệ.

Nhiều tập đoàn công nghệ lớn như Google, Facebook, Twitter, Microsoft và TikTok đã đưa ra các giải pháp nhằm hạn chế sự lan truyền tin giả:

Năm 2016, Facebook hợp tác với các tổ chức thuộc Mạng lưới Kiểm chứng Quốc tế (IFCN) để xác minh và đánh dấu các bài viết chứa thông tin sai lệch [3].

Google triển khai sáng kiến Google News Initiative nhằm giảm thiểu sự lan truyền tin giả và nâng cao chất lượng tin tức trên nền tảng của mình [4].

Twitter áp dụng các chính sách kiểm soát nội dung chặt chẽ hơn, bao gồm dán nhãn cảnh báo và hạn chế phạm vi tiếp cận của các bài đăng không chính xác [5].

Microsoft phát triển Project Origin, một dự án nhằm xác thực tính nguyên bản của nội dung số [6].

TikTok thiết lập hệ thống kiểm chứng nội dung và hợp tác với các tổ chức kiểm tra thông tin để hạn chế sự phát tán tin giả trên nền tảng của mình [7].

Ngoài ra, các bài đăng trên mạng xã hội ngày nay không chỉ chứa văn bản mà còn bao gồm hình ảnh và video, tạo ra những thách thức mới trong việc phát hiện tin giả. Điều này thúc đẩy sự phát triển của các phương pháp nhận diện tin giả dựa trên nội dung đa phương tiện. Vì vậy, nghiên cứu về phát hiện tin giả đang trở thành một chủ đề quan trọng, thu hút sự quan tâm rộng rãi từ giới học thuật và ngành công nghệ trên toàn cầu.

2. TIN TỨC GIẢ, CÁC THUẬT NGỮ LIÊN QUAN VÀ CÔNG CỤ XÁC MINH

Theo Từ điển Cambridge, tin tức giả (Fake News) được định nghĩa là "những câu chuyện sai sự thật được tạo ra và lan truyền trên Internet nhằm tác động đến dư luận và khiến chúng có vẻ đáng tin" [8].

Các thuật ngữ liên quan đến tin giả:

Tuyên truyền (Propaganda): Thông tin được tạo ra và lan truyền bởi một tổ chức chính trị nhằm tác động đến quan điểm của công chúng.

Thông tin sai lệch (Misinformation): Thông tin không chính xác, có thể được lan truyền do nhầm lẫn hoặc vô ý mà không có ý định đánh lừa người khác.

Thông tin đánh lừa (Disinformation): Thông tin sai lệch được cố ý tạo ra và lan truyền với mục đích thao túng sự thật và gây hiểu lầm cho công chúng.

Tin đồn và lừa đảo (Rumors & Hoaxes): Thông tin bịa đặt hoặc bị xuyên tạc có chủ đích, thường được trình bày như thể chúng đã được kiểm chứng bởi các phương tiện truyền thông chính thống.

Châm biếm và hài hước (Parody & Satire): Sử dụng yếu tố hài hước để cung cấp thông tin, mô phỏng phong cách của các hãng tin tức lớn, nhưng người đọc có thể

hiểu nhầm những nội dung này là tin thật.

Tiêu đề giật gân (Clickbait): Tiêu đề gây sốc hoặc hấp dẫn nhằm thu hút sự chú ý của người dùng và khuyến khích họ nhấp vào bài viết, thường để tăng doanh thu quảng cáo [8].

Các công cụ kiểm chứng tin giả:

Tổng hợp các công cụ phát hiện tin giả và đặc điểm của chúng được thể hiện ở Bảng 1.

Bảng 1. Tổng hợp các công cụ phát hiện tin giả và đặc điểm của chúng

Tên công cụ	Đặc điểm chính	Phương pháp sử dụng	Ưu điểm	Hạn chế
Snopes	Một trong những nền tảng kiểm chứng thông tin lâu đời nhất	Kiểm chứng thủ công bởi chuyên gia	Độ tin cậy cao, có cơ sở dữ liệu lớn	Chậm, khó mở rộng
PolitiFact	Đánh giá tính chính xác của các tuyên bố chính trị	Hệ thống "Truth-O-Meter", kiểm chứng thủ công	Độ chính xác cao trong lĩnh vực chính trị	Chỉ tập trung vào chính trị
FactCheck.org	Kiểm chứng tin tức liên quan đến chính trị, khoa học	Kiểm chứng bởi tổ chức phi lợi nhuận	Độ tin cậy cao, có tài liệu tham khảo chi tiết	Không bao phủ nhiều lĩnh vực khác
Hoaxy	Phân tích cách tin tức lan truyền trên mạng xã hội	Thuật toán phân tích mạng xã hội	Xác định mô hình lan truyền tin giả	Không đánh giá chính xác nội dung tin tức
NewsGuard	Đánh giá độ tin cậy của các trang tin tức	Đánh giá trang web dựa trên tiêu chí minh bạch	Đánh giá rõ ràng về nguồn tin	Không kiểm chứng từng bài báo cụ thể
TweetCred	Đánh giá độ tin cậy của tweet trên Twitter	Mô hình học máy, đánh giá theo thời gian thực	Phát hiện tin giả nhanh trên Twitter	Giới hạn trong nền tảng Twitter
Google Fact Check Explorer	Công cụ của Google tổng hợp kết quả kiểm chứng từ nhiều nguồn	Tổng hợp từ nhiều tổ chức kiểm chứng	Phạm vi kiểm chứng rộng, tích hợp AI	Phụ thuộc vào nguồn kiểm chứng có sẵn

Media Bias/Fact Check	Phân loại nguồn tin theo độ thiên vị và độ tin cậy	Đánh giá thủ công và tự động	Đánh giá cả tính khách quan và độ tin cậy	Có thể có sai sót chủ quan
Fake News Detector (AI - based)	Công cụ phát hiện tin giả tự động sử dụng AI	Mô hình học sâu NLP phân tích văn bản	Khả năng phát hiện tin giả nhanh	Độ chính xác phụ thuộc vào dữ liệu huấn luyện
Full Fact AI	Kiểm chứng tự động dựa trên AI	Xử lý ngôn ngữ tự nhiên (NLP)	Khả năng mở rộng cao	Vẫn cần con người can thiệp để đảm bảo độ chính xác

3. CÁC PHƯƠNG PHÁP PHÁT HIỆN TIN GIẢ

Các phương pháp truyền thống thường dựa vào các đặc điểm được xây dựng thủ công. Tuy nhiên, với sự phát triển của dữ liệu lớn (big data) và sự gia tăng mạnh mẽ của dữ liệu do người dùng tạo ra, các phương pháp hiện đại đang chuyển sang khai thác các đặc trưng sâu hơn (deep-level features). Trong phần này, chúng tôi trình bày các nghiên cứu tiên tiến về phát hiện tin giả, phân loại theo hai nhóm chính:

- Dựa trên nội dung tin tức (Content-based methods).

- Dựa trên bối cảnh xã hội của tin tức (Social context-based methods).

3.1. Phương pháp dựa trên nội dung

Phương pháp này tập trung vào việc phân tích nội dung của bài báo, bao gồm văn bản, hình ảnh hoặc cả hai, để phát hiện tin giả. Các nghiên cứu trong lĩnh vực này chủ yếu dựa trên hai loại đặc trưng:

- Đặc trưng ẩn (latent features).

- Đặc trưng được xây dựng thủ công (hand-crafted features).

Các phương pháp dựa trên nội dung bao gồm bốn nhóm chính:

3.1.1. Dựa trên cơ sở tri thức

Phương pháp này sử dụng kỹ thuật kiểm chứng thông tin (fact-checking) [9], trong đó nội dung tin tức được so sánh với các nguồn dữ liệu bên ngoài để xác minh tính chính xác. Kiểm chứng thông tin có thể chia thành hai loại:

- **Kiểm chứng thủ công (Manual Fact-Checking):**

Dựa vào chuyên gia (Expert-Based Fact-Checking): Các chuyên gia trong lĩnh vực sẽ kiểm tra tính chính xác của thông tin. Một số trang web nổi tiếng sử dụng phương pháp này gồm Snopes, PolitiFact, GossipCop. Dù có độ tin cậy cao, nhưng phương pháp này tốn nhiều thời gian và khó mở rộng khi phải xử lý lượng thông tin khổng lồ trên mạng xã hội [9], [10].

Dựa vào cộng đồng (Crowdsourced Fact-Checking): Sử dụng "trí tuệ đám đông" để kiểm chứng thông tin. Tuy nhiên, phương pháp này dễ bị sai lệch, có thể có nhiều ý kiến trái chiều và độ tin cậy thấp hơn so với chuyên gia [11].

- **Kiểm chứng tự động (Automatic Fact-Checking):**

Do các phương pháp thủ công gặp nhiều hạn chế khi xử lý lượng dữ liệu lớn trên mạng xã hội, các phương pháp kiểm chứng tự động đã được phát triển [12].

Quy trình kiểm chứng tự động thường gồm hai giai đoạn:

- Trích xuất dữ kiện (Fact Extraction): Thu thập thông tin và xây dựng cơ sở tri thức.

- Xác minh thông tin (Fact-Checking): So sánh thông tin với cơ sở tri thức để đánh giá tính chính xác.

Mặc dù có nhiều ưu điểm, nhưng các phương pháp này vẫn gặp thách thức do dữ liệu thực tế thường không đầy đủ, không có cấu trúc và chứa nhiều nhiễu [10].

3.1.2. Dựa trên phong cách viết

Phương pháp này phân tích phong cách viết của bài báo để xác định liệu tác giả có ý định đánh lừa người đọc hay không. Những bài viết giật gân, gây sốc thường có đặc điểm như: Tiêu đề chứa nhiều chữ IN HOA, sử dụng nhiều danh từ riêng, có ít từ dừng (stop words).

Phương pháp này sử dụng phân tích phong cách viết để phát hiện tin giả siêu phân cực và phát hiện sự lừa dối trong văn bản.

3.1.3. Dựa trên ngôn ngữ

Phương pháp này phân tích văn bản ở nhiều mức độ khác nhau: Từ vựng (Lexicon-Level), cú pháp (Syntax-Level), ngữ nghĩa (Semantic-Level), cấu trúc văn bản (Discourse-Level).

Nhược điểm của phương pháp này là khó áp dụng cho các bài đăng ngắn trên mạng xã hội và không thể xử lý tin giả chỉ có hình ảnh hoặc video.

3.1.4. Dựa trên nội dung hình ảnh

Hình ảnh và video có thể tăng mức độ tin cậy của tin giả, khiến người đọc dễ dàng tin vào nội dung được lan truyền mà không kiểm chứng. Điều này đặc biệt nguy hiểm khi các hình ảnh bị chỉnh sửa, cắt ghép hoặc sử dụng ngoài ngữ cảnh để tạo ra thông tin sai lệch.

Các phương pháp phát hiện tin tức giả được chúng tôi so sánh ở Bảng 2.

Bảng 2. So sánh các phương pháp phát hiện tin tức giả

Tiêu chí	Phương pháp dựa trên nội dung	Phương pháp dựa trên bối cảnh xã hội	Phương pháp kết hợp
Cách tiếp cận	Phân tích văn bản, hình ảnh, video để xác định tính xác thực	Xem xét cách tin tức lan truyền trên mạng xã hội	Kết hợp cả nội dung và bối cảnh xã hội
Ưu điểm	- Có thể phát hiện tin giả ngay khi xuất hiện - Áp dụng được cho nhiều định dạng (văn bản, hình ảnh, video)	- Cung cấp thông tin bổ sung giúp đánh giá tin tức chính xác hơn - Phát hiện tài khoản, mạng lưới lan truyền tin giả	- Tận dụng được ưu điểm của cả hai phương pháp - Cải thiện độ chính xác

Nhược điểm	- Dễ bị đánh lừa nếu tin giả có nội dung tương tự tin thật - Khó xử lý tin giả có hình ảnh hoặc video bị chỉnh sửa	- Không phát hiện được tin giả chưa lan truyền rộng rãi - Cần dữ liệu lớn từ mạng xã hội	- Yêu cầu tài nguyên tính toán cao hơn - Cần tích hợp nhiều loại dữ liệu khác nhau
Công nghệ sử dụng	Xử lý ngôn ngữ tự nhiên (NLP), thị giác máy tính (Computer Vision)	Phân tích mạng xã hội, mô hình lan truyền thông tin	Học sâu đa mô thức (Multimodal Deep Learning)
Ứng dụng phổ biến	- Phát hiện tin giả dựa trên nội dung bài báo - Phát hiện ảnh hoặc video bị chỉnh sửa	- Phân tích hành vi người dùng trên mạng xã hội - Đánh giá độ tin cậy của nguồn tin	- Kết hợp kiểm chứng nội dung và theo dõi lan truyền tin tức trên mạng xã hội

3.2. Phương pháp dựa trên bối cảnh xã hội

Các phương pháp phát hiện tin giả dựa trên bối cảnh xã hội có thể chia thành ba hướng chính:

3.2.1. Phương pháp dựa trên mạng lưới xã hội

Các bài đăng trên mạng xã hội thường không tồn tại độc lập mà được lan truyền thông qua các tương tác như chia sẻ, bình luận và gắn thẻ. Do đó, phân tích mối quan hệ giữa người dùng có thể giúp phát hiện tin giả hiệu quả [13]. Một số kỹ thuật quan trọng bao gồm:

- **Phân tích đồ thị lan truyền tin tức:** Mỗi bài đăng trên mạng xã hội có thể được mô hình hóa thành một cây lan truyền (Propagation Tree) hoặc một mạng đồ thị (Graph Network), trong đó các nút là người dùng và các cạnh thể hiện mối quan hệ chia sẻ thông tin.

Xây dựng mô hình dựa trên học máy: Các mô hình AI có thể học từ dữ liệu mạng xã hội để phân biệt tin thật và tin giả. Ví dụ, mô hình **CSI (Crowd-Sourced Intelligence)** sử dụng sự kết hợp giữa đặc trưng nội dung, hành vi người dùng và cấu trúc mạng xã hội để xác định mức độ tin cậy của tin tức.

Phân tích vai trò của các tài khoản lan truyền tin tức: Một số tài khoản có hành vi đáng ngờ như chỉ đăng tin tức từ một nguồn duy nhất, đăng bài với tần suất cao bất thường hoặc chia sẻ nội dung giật gân để thu hút sự chú ý. Các công cụ phát hiện bot và tài khoản

- giả mạo có thể giúp xác định các tài khoản tham gia lan truyền tin giả.

3.2.2. Phương pháp dựa trên thời gian

Thời gian xuất hiện và sự thay đổi nội dung của tin tức trên mạng xã hội có thể cung cấp nhiều thông tin quan trọng để phát hiện tin giả. Một số kỹ thuật chính bao gồm:

- **Phân tích lịch sử chỉnh sửa và cập nhật tin tức:** Một số tin giả có thể xuất hiện nhiều lần dưới các tiêu đề hoặc hình thức khác nhau, đặc biệt là trong các chiến dịch thông tin sai lệch kéo dài. Việc theo dõi lịch sử chỉnh sửa của các bài báo và bài đăng trên mạng xã hội có thể giúp phát hiện những thay đổi bất thường.
- **Xác định mô hình lan truyền theo**

thời gian: Tin giả thường có xu hướng lan truyền nhanh chóng trong một khoảng thời gian ngắn, trong khi tin thật có tốc độ lan truyền ổn định hơn. Tin giả thường đạt mức đỉnh lan truyền nhanh hơn nhưng cũng bị mất tương tác nhanh chóng so với tin thật.

- **Phân tích tương tác theo thời gian:** Việc theo dõi số lượt thích, bình luận, chia sẻ theo thời gian có thể giúp phát hiện các chiến dịch phát tán tin giả được tổ chức bài bản, trong đó một lượng lớn tài khoản giả mạo có thể tương tác với bài viết trong thời gian ngắn để tạo hiệu ứng lan truyền mạnh.

3.2.3. Phương pháp dựa trên độ tin cậy

Một trong những cách hiệu quả để phát hiện tin giả là đánh giá mức độ tin cậy của nguồn tin và người lan truyền thông tin. Một số chiến lược quan trọng bao gồm:

- **Xác minh nguồn gốc tin tức:** Tin giả thường xuất phát từ các trang web không rõ nguồn gốc hoặc có lịch sử phát tán thông tin sai lệch. Việc kiểm tra uy tín của trang web, tên miền, và tác giả bài viết có thể giúp đánh giá tính xác thực của tin tức. Các nền tảng như **NewsGuard** và **Media Bias/Fact Check** cung cấp thông tin chi tiết về mức độ tin cậy của các trang tin tức.
- **Phân tích hồ sơ và hành vi người dùng:** Những tài khoản chuyên lan truyền tin giả thường có đặc điểm giống nhau, chẳng hạn như số lượng bài đăng lớn, nội dung mang tính thiên vị, hoặc có nhiều bài viết bị gắn cờ vi phạm tiêu chuẩn cộng đồng. Các hệ thống đánh giá uy tín của người dùng có thể giúp xác định mức độ đáng tin cậy của người đăng bài.
- **Hệ thống chấm điểm tin cậy (Credibility Scoring):** Một số mô hình AI đã được phát triển để đánh giá mức độ tin cậy của tin tức dựa trên nhiều tiêu chí như nguồn tin, nội dung, cách thức lan truyền và phản hồi từ người đọc. Ví dụ, công cụ **TweetCred** có thể đánh giá mức độ đáng tin cậy của tweet theo thời gian thực dựa trên các đặc điểm như nội dung, tác giả và mức độ lan truyền.

3.3. Kết hợp các phương pháp để nâng cao hiệu quả

Mặc dù các phương pháp dựa trên nội dung và bối cảnh xã hội có thể giúp phát hiện tin giả một cách độc lập, nhưng việc kết hợp chúng lại có thể nâng cao độ chính xác và hiệu quả. Một số hướng kết hợp phổ biến bao gồm:

Kết hợp phân tích nội dung và mạng xã hội: Các mô hình AI có thể tận dụng thông tin từ cả nội dung bài viết và dữ liệu mạng xã hội để đưa ra quyết định tốt hơn. Ví dụ, mô hình **SAFE (Social-Aware Fake News Detection)** sử dụng cả đặc trưng ngôn ngữ và thông tin lan truyền để phát hiện tin giả [14].

- **Ứng dụng học sâu (Deep Learning) để tối ưu hóa phát hiện tin giả:** Các mô hình học sâu có thể tự động học từ dữ liệu lớn và cải thiện độ chính xác trong việc phát hiện tin giả. Các kỹ thuật như **Transformer-based models (BERT, RoBERTa, GPT)** có thể giúp phân tích ngữ nghĩa nâng cao, trong khi **Graph Neural Networks (GNNs)** có thể giúp phát hiện tin giả dựa trên cấu trúc mạng xã hội.

4. CÁC TẬP DỮ LIỆU DÙNG ĐỂ PHÁT HIỆN TIN GIẢ

Nhiều nhà nghiên cứu đã xây dựng các tập dữ liệu chứa tin giả để phục vụ việc huấn luyện mô hình. Tuy nhiên, chỉ một số tập dữ liệu chuẩn mực (benchmark datasets) được công bố công khai.

Một số tập dữ liệu phổ biến được sử dụng để phát triển các mô hình và thuật toán phát hiện tin giả được thể hiện ở Bảng 3.

Bảng 3. So sánh một số tập dữ liệu phổ biến dùng để phát hiện tin giả

Tên tập dữ liệu	Nguồn dữ liệu	Ngôn ngữ	Số lượng mẫu	Loại tin giả	Đặc điểm nổi bật	Hạn chế
LIAR	Các tuyên bố chính trị từ Politifact	Tiếng Anh	~12,800	Tin tức sai lệch, bóp méo sự thật	Nhãn chi tiết với 6 mức độ tin cậy	Chủ yếu tập trung vào chính trị
FakeNewsNet	BuzzFeed, PolitiFact	Tiếng Anh	~23,500	Tin tức sai lệch trên mạng xã hội	Chứa cả nội dung bài báo và thông tin lan truyền	Dữ liệu mạng xã hội hạn chế
BuzzFeed News	Facebook	Tiếng Anh	1,500	Tin tức sai lệch trên mạng xã hội	Được dán nhãn bởi chuyên gia	Quy mô nhỏ, chỉ có dữ liệu Facebook
FEVER (Fact Extraction and Verification)	Wikipedia	Tiếng Anh	~185,000	Tin sai lệch dựa trên sự kiện thực tế	Mức độ chi tiết cao, có nguồn tham chiếu	Chủ yếu là tin sai lệch do chỉnh sửa thông tin
Covid-19 Fake News Dataset	Twitter, Fact-checking websites	Tiếng Anh	~10,000	Tin giả về Covid-19	Tập trung vào một chủ đề quan trọng	Hạn chế về phạm vi chủ đề
GossipCop	GossipCop	Tiếng Anh	~50,000	Tin đồn, tin giả giải trí	Phân loại tin đồn chính xác	Chủ yếu tập trung vào tin giải trí
Twitter Fake News	Twitter	Tiếng Anh	~17,000	Tin giả lan truyền trên Twitter	Chứa thông tin về tài khoản và retweet	Giới hạn trong nền tảng Twitter

Các tập dữ liệu này rất hữu ích cho việc huấn luyện và đánh giá các mô hình học máy, đặc biệt trong các bài toán phân loại văn bản để phát hiện tin giả.

5. TỔNG QUAN CÁC NGHIÊN CỨU TRONG LĨNH VỰC PHÁT HIỆN TIN GIẢ

Trong lĩnh vực phát hiện tin giả, có nhiều vấn đề và thách thức đáng chú ý mà các nhà nghiên cứu và chuyên gia đang phải đối mặt. Dưới đây là một số vấn đề chính và thách thức quan trọng:

5.1. Chất lượng dữ liệu

- **Dữ liệu không đồng nhất, đầy đủ**: Các tập dữ liệu dùng để huấn luyện mô hình phát hiện tin giả thường không đồng nhất về cấu trúc, nội dung và độ chính xác. Tập dữ liệu không phải lúc nào cũng bao phủ được đủ các dạng tin giả, đặc biệt là những tin giả có nội dung phức tạp hoặc tin tức trong những bối cảnh đặc biệt. Điều này dẫn đến việc mô hình học máy dễ bị thiếu sót và không thể nhận diện được tất cả các kiểu tin giả.
- **Nhãn dữ liệu không chính xác**: Việc xác định nhãn "thật" hoặc "giả" trong dữ liệu là rất quan trọng. Tuy nhiên, trong một số trường hợp, nhãn có thể không chính xác do sự thiên vị của người gán nhãn hoặc thiếu sự thống nhất trong cách phân loại tin tức.

5.2. Đặc điểm ngữ nghĩa và biến thể của tin giả

Tin giả có thể thay đổi cách thức và hình thức truyền tải tùy theo ngữ cảnh, ví dụ như cách diễn đạt, ngữ điệu, hoặc nguồn gốc. Việc phát hiện các biến thể ngữ nghĩa của tin giả là một thách thức lớn. Tin giả có thể được chỉnh sửa hoặc viết lại nhiều lần để đánh lừa người đọc. Những thay đổi này có thể làm cho tin giả trông giống như một bài báo chính thức hoặc có vẻ hợp lý hơn, khiến việc nhận diện trở nên phức tạp hơn.

5.3. Xử lý các nguồn tin không tin cậy

Việc kiểm tra nguồn gốc của tin tức không phải lúc nào cũng dễ dàng, đặc biệt là khi nguồn tin có thể là các blog không đáng tin cậy, tài khoản mạng xã hội giả mạo, hoặc các trang web không xác minh được tính xác thực. Đôi khi, một tin tức sai sự thật không phải là một tin giả có chủ đích, mà là một tin không chính xác do sai sót trong quá trình báo cáo. Phân biệt rõ ràng giữa tin giả và tin sai sót là rất quan trọng để tránh gây nhầm lẫn.

5.4. Sự phát triển của tin giả

Các công nghệ như deepfake cho phép tạo ra video, hình ảnh hoặc âm thanh giả mạo, khiến việc phát hiện tin giả trở nên ngày càng khó khăn hơn. Với sự phát triển của các mô hình ngôn ngữ tự nhiên tiên tiến như GPT, AI có thể tạo ra các bài viết giả mạo có nội dung hoàn chỉnh và đáng tin cậy. Điều này tạo ra một thách thức lớn cho các hệ thống phát hiện tin giả khi các tin giả có thể được tạo ra tự động và không có dấu hiệu dễ nhận biết.

5.5. Khó khăn trong việc phân loại tin giả

Các mô hình học máy có thể bị thiên vị khi phân loại tin tức, đặc biệt là khi tập dữ liệu huấn luyện không đa dạng. Một số tin giả có thể được thiết kế rất tinh vi, có nội dung không rõ ràng hoặc có thể gây khó khăn trong việc nhận diện. Những tin giả này thường không có yếu tố gây nghi ngờ rõ ràng, khiến mô hình phân loại dễ bị nhầm lẫn.

5.6. Vấn đề mạng xã hội và tốc độ lây lan

Tin giả có thể lan truyền rất nhanh chóng trên các nền tảng mạng xã hội, như Facebook, Twitter, hoặc TikTok. Điều này tạo ra một vấn đề lớn cho việc phát hiện và

ngăn chặn tin giả kịp thời. Một khi tin giả được lan truyền rộng rãi, việc kiểm tra và xác minh lại thông tin có thể rất khó khăn. Ngoài ra, sự xuất hiện của các tài khoản tự động (bots) và những người sử dụng mạng xã hội để thao túng dư luận (trolls) có thể phát tán tin giả với tốc độ rất nhanh. Việc nhận diện và ngăn chặn các tài khoản này là một thách thức lớn đối với các nền tảng và hệ thống phát hiện tin giả.

5.7. Vấn đề liên quan đến bảo mật và quyền riêng tư

Việc thu thập dữ liệu từ mạng xã hội hoặc các nguồn tin khác để huấn luyện mô hình phát hiện tin giả có thể vi phạm quyền riêng tư của người dùng nếu không được thực hiện một cách thận trọng. Các hệ thống phát hiện tin giả cần phải đảm bảo rằng việc sử dụng công nghệ không gây ra những lỗ hổng bảo mật có thể bị khai thác bởi các đối tượng xấu.

5.8. Sự thiếu sự thống nhất trong các quy định và tiêu chuẩn

Mặc dù có nhiều mô hình và kỹ thuật được sử dụng để phát hiện tin giả, nhưng vẫn chưa có một tiêu chuẩn chung hoặc cách tiếp cận thống nhất trong ngành. Các nền tảng mạng xã hội, tổ chức kiểm tra sự thật và các nhà nghiên cứu thường làm việc riêng biệt mà không có sự hợp tác đồng bộ. Điều này làm giảm hiệu quả trong việc phát hiện và ngăn chặn tin giả trên quy mô toàn cầu.

6. KẾT LUẬN

Sự phát triển mạnh mẽ của mạng xã hội trong những năm gần đây đã thay đổi cách con người tiếp cận và tiêu thụ tin tức. Ngày càng có nhiều độc giả lựa chọn cập nhật thông tin qua các nền tảng trực tuyến thay vì các kênh truyền thống. Tuy nhiên, điều này cũng tạo điều kiện thuận lợi cho việc lan truyền tin giả và thông tin sai lệch, gây ảnh hưởng đến nhận thức của công chúng. Bài báo này cung cấp cái nhìn tổng quan về các phương pháp phát hiện tin giả. Ngoài ra, bài viết cũng giới thiệu các tập dữ liệu phổ biến và các công cụ hỗ trợ kiểm chứng thông tin.

Mặc dù các phương pháp phát hiện tin giả đã đạt được nhiều tiến bộ, nhưng vẫn còn nhiều thách thức cần được giải quyết, bao gồm: Phát hiện tin giả theo thời gian thực, tích hợp nhiều nguồn dữ liệu hơn, nâng cao tính minh bạch của mô hình, xây dựng hệ thống tự động phản bác tin giả…

Việc tiếp tục nghiên cứu và phát triển các công cụ phát hiện tin giả hiệu quả sẽ giúp người dùng tiếp cận thông tin chính xác hơn, đồng thời góp phần bảo vệ sự thật và hạn chế tác động tiêu cực của tin giả đối với xã hội.

TÀI LIỆU THAM KHẢO

[1] Allcott, Hunt, and Matthew Gentzkow. “Social media and fake news in the 2016 election”. *Journal of economic perspectives* 31.2 (2017): 211-236.

[2] Apuke, Oberiri Destiny, and Bahiyah Omar. “Fake news and COVID-19: modelling the predictors of fake news sharing among social media users”. *Telematics and informatics* 56 (2021): 101475.

[3] Figueira, Álvaro, and Luciana Oliveira. “The current state of fake news: challenges and opportunities”. *Procedia computer science* 121 (2017): 817-825.

[4] Burkhardt, Joanna M. *Combating fake news in the digital age*. Vol. 53. No. 8.

Chicago, IL, USA:: American Library Association, 2017.

[5] Khalil, Ashraf, Hassan Hajjdiab, and Nabeel Al-Qirim. "Detecting fake followers in twitter: A machine learning approach". *International Journal of Machine Learning and Computing* 7.6 (2017): 198-202.

[6] Aythora, J., et al. "Multi-stakeholder media provenance management to counter synthetic media risks in news publishing"." *Proc. Intl. Broadcasting Convention (IBC)*. Vol. 1. No. 2. 2020.

[7] Lan, Duong Hoai, and Tran Minh Tung. "Exploring fake news awareness and trust in the age of social media among university student TikTok users". *Cogent Social Sciences* 10.1 (2024): 2302216.

[8] Zhou, Xinyi, and Reza Zafarani. "A survey of fake news: Fundamental theories, detection methods, and opportunities". *ACM Computing Surveys (CSUR)* 53.5 (2020): 1-40.

[9] Pilarski, Moritz, Kirill Olegovich Solovev, and Nicolas Pröllochs. "Community notes vs. snoping: how the crowd selects fact-checking targets on social media". *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 18. 2024.

[10] Vo, Nguyen, and Kyumin Lee. "The rise of guardians: Fact-checking url recommendation to combat fake news". *The 41st international ACM SIGIR conference on research & development in information retrieval*. 2018.

[11] Capuano, Nicola, et al. "Content-based fake news detection with machine and deep learning: A systematic review". *Neurocomputing* 530 (2023): 91-103.

[12] Pérez-Rosas, Verónica, et al. "Automatic detection of fake news". *arXiv preprint arXiv:1708.07104* (2017).

[13] Kumar, Akshi, and Geetanjali Garg. "Systematic literature review on context-based sentiment analysis in social multimedia". *Multimedia tools and Applications* 79.21 (2020): 15349-15380.

[14] Peng, Kai, et al. "TOFDS: A two-stage task execution method for fake news in digital twin-empowered socio-cyber world". *IEEE Transactions on Computational Social Systems* (2023).

Liên hệ:

TS. Hoàng Đình Tuyền

Khoa Công nghệ thông tin, Trường Đại học Quảng Bình

Địa chỉ: 18 Nguyễn Văn Linh, Đồng Hới, Quảng Bình

Email: hoangdinhtuyen@gmail.com

Ngày nhận bài: 13/02/2025

Ngày gửi phản biện: 13/02/2025

Ngày duyệt đăng: 26/02/2025