PHÁT HIỆN TỰ ĐỘNG TIN GIẢ: THÀNH TỰU VÀ THÁCH THỨC

AUTOMATIC FAKE NEWS DETECTION: ACHIEVEMENTS AND CHALLENGES

Võ Trung Hùng[1*], Ninh Khánh Chi[2], Trần Anh Kiệt[3]

[1]Trường Đại học Sư phạm Kỹ thuật - Đại học Đà Nẵng
[2]Trường Đại học CNTT & Truyền thông Việt-Hàn - Đại học Đà Nẵng
[3]Đại học Đà Nẵng

*Tác giả liên hệ: vthung@ute.udn.vn

(Nhận bài: 06/01/2022; Chấp nhận đăng: 27/02/2022)

Tóm tắt - Trong bài báo này, nhóm tác giả trình bày một cách tổng quan các vấn đề liên quan đến khái niệm, phân loại, cách xác định thủ công và xác định tự động các tin giả. Đặc biệt, nhóm tác giả đã trình bày hai kỹ thuật được ứng dụng rộng rãi hiện nay đó là kỹ thuật học máy và kỹ thuật học sâu. Hai kỹ thuật này đều dựa trên phân tích nội dung bản tin và bước đầu đã mang lại những kết quả tích cực. Tuy nhiên, đây là bài báo mang tính chất nghiên cứu tổng quan nên nhóm tác giả chỉ dừng ở mức tổng hợp, phân tích, nhận định và trình bày lại những kết quả nghiên cứu đã có trước đó. Đóng góp chính trong bài báo này là chỉ ra được những thách thức và hướng nghiên cứu sắp đến cho tiếng Việt trong lĩnh vực phát hiện tin giả.

Từ khóa - Tin giả; phát hiện tự động; mạng nơ-ron; học máy; học sâu

Abstract - In this paper, the authors present an overview of issues related to the concept, classification, manual detection and automatic detection of fake news. In particular, the authors present two widely applied techniques today: Traditional machine learning and deep learning. These two techniques are based on content analysis and initially offered positive results. However, this article is of an overview research, therefore we only stop at the level of synthesizing, analyzing, commenting and presenting previous research results. Our main contribution in this paper is to point out the challenges and upcoming research directions for Vietnamese in the field of fake news detection.

Key words - Fake news; automatic detection; neural networks; machine learning; deep learning

1. Đặt vấn đề

Thời gian qua, trên mạng Internet, đặc biệt là các trang mạng xã hội, xuất hiện một số tài khoản giả mạo, đăng các thông tin không kiểm chứng liên quan đến nhiều chủ đề về chính trị, dịch bệnh, thiên tai, khí tượng thủy văn, mê tín dị đoan, quảng cáo sai sự thật... Việc này gây hoang mang, xáo trộn, ảnh hưởng lớn đến đời sống sinh hoạt của người dân.

Để chấn chỉnh tình trạng trên, Bộ Thông tin và Truyền thông đã khai trương cổng thông tin tiếp nhận phản ánh, công bố tin giả, ra mắt đầu số tiếp nhận phản ánh tin giả. Theo đó, cổng thông tin tiếp nhận thông báo tin giả có địa chỉ tên miền www.tingia.gov.vn, do Trung tâm Xử lý tin giả Việt Nam (VAFC) trực thuộc Cục Phát thanh - Truyền hình và thông tin điện tử của Bộ Thông tin và Truyền thông quản lý. Các nhiệm vụ, chức năng của Trung tâm xử lý tin giả Việt Nam gồm: Phối hợp các cơ quan chức năng để thẩm định, công bố tin giả; Đánh giá xu hướng thông tin chia sẻ, tương tác lớn để dán nhãn cảnh báo tin giả; Tiếp nhận, phát hiện, thẩm định, gắn nhãn tin giả; Công bố thông tin xác thực; Hướng dẫn cách nhận biết, phòng tránh, đối phó với tin giả. Trung tâm tập trung vào các lĩnh vực thông tin như sau: Chính sách, pháp luật; Kinh tế, tài chính; Lĩnh vực y tế, sản phẩm y tế liên quan đến sức khỏe con người; Thiên tai, dịch bệnh; An ninh quốc gia, trật tự an toàn – xã hội; Tài khoản giả mạo; Đường link lừa đảo; Các lĩnh vực khác (https://tingia.gov.vn).

Tuy nhiên, việc phát hiện tin giả ở Việt Nam hiện nay được thực hiện hoàn toàn thủ công bởi con người, như triết lý hành động được công bố của VAFC *"Tin giả do con người tạo ra nên chỉ có duy nhất con người mới có thể nhận biết và xử lý được tin giả"*.

Có rất ít các nghiên cứu về phát hiện tin giả được công bố bởi các tác giả trong nước. Bài báo [1] trình bày cách tiếp cận phát hiện tin tức giả mạo trên các trang web mạng xã hội (SNS - Social Network Sites), bằng phương pháp tổng hợp các đặc điểm ngôn ngữ được sử dụng PhoBERT. Bài báo [2] trình bày về những nhiệm vụ được chia sẻ trên ReINTEL bao gồm ba giai đoạn: Khởi động, thử nghiệm công khai, thử nghiệm riêng tư liên quan đến tin giả. Những kết quả này còn hết sức sơ khai và chưa thể áp dụng.

Trên thế giới, các nghiên cứu hiện tại thường kết nối tin tức giả mạo với các thuật ngữ và khái niệm như tin tức lừa đảo [3], [4]. Những thách thức của nghiên cứu tin tức giả bắt đầu từ việc xác định thế nào là tin tức giả. Cho đến nay, không có định nghĩa chung nào được cung cấp cho tin tức giả mạo, nơi nó được coi là "một bài báo sai sự thật có chủ ý và khó có thể xác minh được" [3], [5].

Các lý thuyết cơ bản về nhận thức và hành vi của con người được phát triển trên nhiều lĩnh vực khác nhau, chẳng hạn như khoa học xã hội và kinh tế, cung cấp những hiểu biết vô giá cho việc phân tích tin tức giả mạo. Những lý thuyết này có thể giới thiệu cơ hội mới cho các nghiên cứu định tính và định lượng về dữ liệu tin tức giả. Những lý thuyết này cũng có thể tạo điều kiện cho việc xây dựng các mô hình hợp lý và có thể giải thích được để phát hiện và can thiệp vào sự phát tán tin tức giả, mà cho đến nay, hiếm khi có sẵn [6]. Các lý thuyết liên quan đến tin tức tiết lộ các đặc điểm có thể có của nội dung tin tức giả mạo so với nội dung

[1] The University of Danang - University of Technology and Education (Trung Hung VO)
[2] The University of Danang - Vietnam-Korea University of Information and Communication Technology (Ninh Khanh Chi)
[3] The University of Danang (Anh Kiet TRAN)

tin tức thật. Ví dụ: Các lý thuyết ngụ ý rằng tin tức giả mạo có khả năng khác với sự thật ở một số điểm như phong cách viết hoặc số liệu thống kê [7] và cách biểu đạt tình cảm [8].

Hiện nay, một số nhà khoa học đã đề xuất ứng dụng các mô hình mạng nơ-ron và học sâu (Deep Learning) cho việc phát hiện tin giả và đã thu được một số kết quả [9], [10].

Gần đây, các nhà nghiên cứu đã kết hợp nhiều mô hình và phương pháp để cải thiện chất lượng của việc phát hiện tin giả. Sự kết hợp của CNN-RNN đã được chứng minh là thành công trong một số nhiệm vụ phân loại và hồi quy, vì chúng có khả năng nắm bắt cả đặc tính cục bộ và tuần tự của dữ liệu đầu vào. Ví dụ, chúng đã được sử dụng để phát hiện cảm xúc [11] hoặc trích chọn các đặc trưng bằng cách kết hợp các mô hình [12].

Tóm lại, so với Việt Nam, trên thế giới đã triển khai nghiên cứu về phát hiện tin giả trong vài năm gần đây và bước đầu đề xuất được một số giải pháp, mô hình mang lại hiệu quả tốt. Tuy nhiên, việc xác định và phát hiện tin giả vẫn còn là một bí ẩn lớn cần khám phá, là một hướng nghiên cứu mới của trí tuệ nhân tạo.

Trong bài báo này, nhóm tác giả trình bày tổng hợp những kết quả đạt được trong lĩnh vực phát hiện tự động tin giả và chỉ ra những thách thức cần phải nghiên cứu giải quyết trong thời gian đến, đặc biệt cho tiếng Việt.

2. Tổng quan về tin giả và phát hiện tin giả

2.1. Khái niệm về tin giả

Thuật ngữ "tin giả" là một khái niệm tương đối mới và cho đến nay vẫn chưa có một định nghĩa chung được thống nhất về tin tức giả mạo hay tin giả (Fake News).

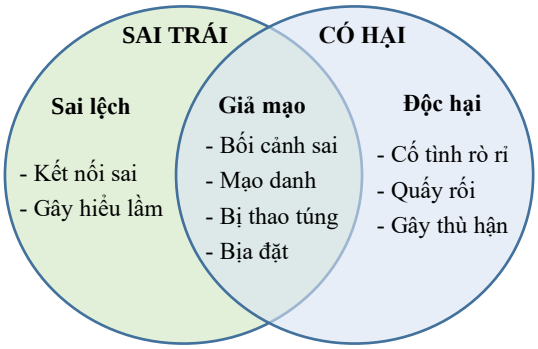
Theo từ điển Oxford *"Tin giả là thông tin sai sự thật được phát sóng hoặc xuất bản dưới dạng tin tức nhằm mục đích lừa đảo hoặc có động cơ chính trị. Tin giả tạo ra sự nhầm lẫn đáng kể của công chúng về các sự kiện hiện tại. Tin giả bùng nổ trên phương tiện truyền thông xã hội, đang xâm nhập vào các kênh truyền thông chính"*.

Học giả về truyền thông Nolan Higdon đã định nghĩa "*Tin tức giả là nội dung sai sự thật hoặc gây hiểu lầm được trình bày dưới dạng tin tức và được truyền đạt dưới các định dạng bao gồm truyền thông nói, viết, in, điện tử và kỹ thuật số"* [13]. Tin tức giả mạo cũng đề cập đến những câu chuyện bịa đặt có rất ít hoặc không có sự thật và khó có thể xác minh được. Thậm chí rộng hơn, sau kỳ bầu cử tổng thống Mỹ năm 2020, người ta đã mở rộng ý nghĩa của "tin tức giả" để bao gồm cả các tin tức tiêu cực về niềm tin và hành động cá nhân của họ.

2.2. Phân loại tin giả

Các trường hợp điển hình của tin giả bao gồm quảng cáo lừa đảo (trong kinh doanh và chính trị), tuyên truyền của chính phủ, các hình ảnh chỉnh sửa hoặc dùng sai mục đích ban đầu, tài liệu giả mạo, bản đồ giả, gian lận trên Internet, các trang web giả mạo và mục từ trên Wikipedia không đúng sự thật,... Tin giả có thể gây ra tác hại đáng kể nếu mọi người để nó lừa dối. Để giải quyết mối đe dọa này đối với chất lượng thông tin, trước tiên chúng ta cần hiểu chính xác các loại tin giả.

Có rất nhiều nghiên cứu về tin giả và phân loại tin giả, một trong những báo cáo được tham khảo và trích dẫn nhiều về phân loại tin giả là của Claire Wardle [14].



Hình 1. *Phân loại tin tức giả mạo*

Theo phân loại này, các tin giả được phân thành 3 nhóm chính (Hình 1):

1) **Thông tin sai lệch (Mis-information)**: Thông tin sai lệch được phổ biến mà không có ý định gây hại. Thông tin sai lệch có 2 loại:

- *Kết nối sai (False connection)*: Khi dòng tiêu đề, hình ảnh hoặc chú thích không phù hợp với nội dung. Ví dụ như trường hợp giật tít để câu view bằng những tiêu đề giật gân nhưng nội dung không phản ánh đúng với tên ở tiêu đề; hoặc sử dụng hình ảnh không đúng với nội dung (chẳng hạn các ảnh rùng rợn hay tươi mát để thu hút người khác truy cập).

- *Nội dung gây hiểu lầm (Misleading content)*: Sử dụng sai thông tin và gây hiểu lầm cho người đọc. Ví dụ, nội dung quảng cáo hoặc trang web cố gắng đánh lừa khách hàng để truy cập vào các trang web không an toàn. Nó có thể bao gồm cả những nội dung có thể được coi là lừa đảo, gian lận hoặc có hại cho khách truy cập trang web một cách hợp lý thông qua các tuyên bố không có căn cứ, ưu đãi miễn phí hoặc hứa hẹn về giảm giá, quảng cáo gây hiểu lầm và quảng bá các sản phẩm và dịch vụ của bên thứ ba.

2) **Thông tin giả mạo (Dis-information)**: Được tạo và chia sẻ bởi những người có ý định gây hại.

- *Bối cảnh sai (False context)*: Loại thông tin giả mạo này được sử dụng để mô tả nội dung xác thực nhưng đã được điều chỉnh lại theo những cách nguy hiểm. Ví dụ, lợi dụng sự cố Formosa xả thải gây ra hiện tượng cá chết hàng loạt tại vùng biển khu vực các tỉnh bắc miền Trung, nhiều bản tin đã lồng ghép các ý đồ chính trị để kích động, chống phá chế độ.

- *Nội dung mạo danh (Imposter content):* Là những nội dung sai sự thật hoặc gây hiểu lầm bằng cách sử dụng các biểu trưng nổi tiếng hoặc tin tức từ các nhân vật hoặc nhà báo có uy tín. Như chúng ta biết, bộ não của con người luôn tìm kiếm từ kinh nghiệm tích luỹ được để xác định độ tin cậy khi tiếp nhận một thông tin nào đó. Dựa trên kinh nghiệm là lối tắt tư duy để giúp chúng ta hiểu được thế giới. Lợi dụng điều này, người tạo tin giả sẽ tìm cách giả mạo là nội dung do những cá nhân, tổ chức nổi tiếng cung cấp hoặc đã được họ chấp nhận. Ví dụ, ở Việt Nam trong thời gian gần đây, các nhãn hàng đã mời các nghệ sĩ nổi tiếng quảng cáo sai sự thật đã trở thành một vấn nạn và gây khó khăn cho sự lựa chọn của khách hàng.

- *Nội dung bị thao túng (Manipulated content)*: Nội dung bị thao túng là khi một khía cạnh nào đó của nội dung chính hãng bị thay đổi. Điều này thường liên quan đến ảnh hoặc video. Ví dụ, khi cố tình đưa tin sai sự thật về một vụ

tai tiếng (scandal) của người nổi tiếng, một thủ thuật thường được sử dụng là ghép ảnh, chỉnh sửa ảnh gốc theo dụng ý của người đưa tin để minh hoạ cho nội dung.

- *Nội dung bịa đặt (Fabricated content)*: Nội dung bịa đặt là sai 100%. Ví dụ, vào tháng 8 năm 2021, trên mạng xã hội chia sẻ với tốc độ chóng mặt về tin một người bác sĩ tên Trần Khoa, người này chia sẻ đã quyết định *"nhường đi chiếc máy thở"* của ba mẹ mình đang dùng cho một sản phụ đang cần. Thông tin này đi kèm với một lá thư rất lâm ly của bác sĩ Khoa và nhận được sự đồng cảm lớn từ cộng đồng mạng. Tuy nhiên, Sở Y tế Thành phố Hồ Chí Minh cho biết sau, khi kiểm tra có đủ cơ sở khẳng định thông tin lan truyền về trường hợp một bác sĩ rút ống thở của người nhà để nhường máy thở cho mẹ con sản phụ là hư cấu.

3) **Thông tin độc hại (Mal-information)**: Chia sẻ thông tin "chính hãng" nhưng với mục đích gây hại.

- *Rò rỉ (Leaks)*: Rò rỉ thông tin là một sự kiện diễn ra khi thông tin bí mật được tiết lộ cho những người hoặc bên không có thẩm quyền. Ví dụ, trong các cuộc bầu cử tổng thống Mỹ hoặc trước các kỳ đại hội Đảng ở Việt Nam thường xuất hiện rất nhiều các thông tin được cho là rò rỉ từ các hồ sơ mật và gần như không thể kiểm chứng. Những thông tin này thường gây hoang mang và tạo ra nhiều luồng dư luận trái chiều.

- *Quấy rối (Harassment)*: Là bất kỳ hành vi nào, dù bằng lời nói, hình ảnh, văn bản hay cách khác nhằm mục đích xúc phạm hoặc làm nhục một cá nhân, tổ chức nào đó. Cùng với mạng xã hội, các hành vi quấy rối ngày càng trở nên phổ biến và tinh vi. Ví dụ, fanpage của các nhân vật nổi tiếng thường lan truyền các thông tin nhằm hạ thấp các đối thủ cạnh tranh và nâng cao hình ảnh thần tượng của mình.

- *Gây chia rẻ, thù hận (Hate speech)*: Những nội dung biểu hiện qua lời nói, văn bản hoặc các biểu hiện khác thể hiện sự căm thù, phỉ báng một người hoặc những người khác. Các nội dung gây chia rẻ, thù hận thường dựa trên một nhóm xã hội được xác định bởi các thuộc tính như chủng tộc, dân tộc, giới tính, khuynh hướng tình dục, tôn giáo, tuổi tác, khuyết tật về thể chất hoặc tinh thần.

2.3. Phát hiện tin giả

Có rất nhiều cách để chúng ta phát hiện tin giả và điều này phụ thuộc vào nhiều yếu tố như kiến thức, kinh nghiệm, kỹ năng phân tích, năng lực phán đoán, tư duy phê phán,... [15].

Dưới đây, nhóm tác giả tổng hợp một số cách thông dụng mà con người sử dụng để phát hiện tin giả trên mạng:

1) Kiểm tra nguồn tin: Kiểm tra địa chỉ web cho trang đang xem hoặc nơi phát tán nội dung. Đôi khi, các trang web tin tức giả mạo có thể có lỗi chính tả trong URL hoặc sử dụng phần mở rộng tên miền ít thông dụng hơn như ".infonet" hoặc ".offer". Việc xác định nguồn gốc phát tán nội dung sẽ giúp đánh giá độ tin cậy của nội dung.

2) Kiểm tra tác giả: Nghiên cứu về tác giả để xem liệu chúng có đáng tin cậy hay không. ví dụ: Tác giả này có thật không, tác giả có danh tiếng tốt không, tác giả có viết về lĩnh vực chuyên môn cụ thể của nội dung phát tán không? Đặc biệt, xem xét động cơ của người viết có thể là gì.

3) Kiểm tra các nguồn khác: Đối chiếu với các cơ quan truyền thông hoặc các tổ chức uy tín khác có đưa tin về câu chuyện này không? Kiểm tra các nguồn đáng tin cậy được trích dẫn trong câu chuyện? Các hãng thông tấn chuyên nghiệp trên toàn cầu có các nguyên tắc biên tập và nhiều nguồn tài nguyên để kiểm tra thực tế, vì vậy nếu họ cũng đang tường thuật câu chuyện, đó là một dấu hiệu tốt.

4) Duy trì tư duy phản biện: Rất nhiều tin tức giả được viết một cách khéo léo để kích động các phản ứng cảm xúc mạnh mẽ như sợ hãi hoặc tức giận để thao túng người đọc. Việc duy trì tư duy phản biện bằng cách tự hỏi bản thân: Tại sao câu chuyện này lại được viết? Nó có đang thúc đẩy một nguyên nhân hoặc chương trình nghị sự cụ thể nào không? Có phải nó đang cố làm cho chúng ta truy cập qua một trang web khác không?

5) Kiểm tra sự thật: Các câu chuyện tin tức đáng tin cậy sẽ bao gồm nhiều dữ kiện như dữ liệu, thống kê, trích dẫn từ các chuyên gia,... Nếu thiếu những thứ này, hãy đặt câu hỏi tại sao. Các báo cáo có thông tin sai lệch thường chứa ngày tháng không chính xác hoặc mốc thời gian bị thay đổi, vì vậy, chúng ta nên kiểm tra thời điểm bài báo được xuất bản và tính lô-gíc của nội dung.

6) Kiểm tra các nhận xét: Ngay cả khi bài báo hoặc video là hợp pháp, các nhận xét bên dưới có thể giúp chúng ta tìm ra sự thật. Lưu ý, các liên kết hoặc nhận xét được đăng để phản hồi nội dung có thể được tự động tạo bởi rô-bốt mạng hoặc những người được thuê để đưa thông tin gây hiểu lầm.

7) Kiểm tra thành kiến của cá nhân: Tất cả chúng ta đều có thành kiến và nên tránh để thành kiến lấn át lý trí khi đánh giá nội dung bài viết. Phương tiện truyền thông xã hội có thể tạo ra các luồng phản hồi bằng cách đề xuất những câu chuyện phù hợp với thói quen duyệt web, sở thích và quan điểm hiện có của cá nhân, cộng đồng. Càng đọc nhiều nguồn và quan điểm đa dạng, chúng ta càng có nhiều khả năng đưa ra kết luận chính xác.

8) Kiểm tra xem đó có phải là một trò đùa hay không: Các trang web châm biếm rất phổ biến và đôi khi không phải lúc nào cũng rõ ràng một câu chuyện chỉ là một trò đùa hay nhại lại. Kiểm tra trang web, tác giả bài viết để xem liệu chúng có nổi tiếng với tác phẩm châm biếm hoặc tạo ra những câu chuyện hài hước hay không để hiểu đúng bản chất của nội dung.

9) Kiểm tra tính xác thực của hình ảnh: Hình ảnh minh họa mà chúng ta thấy đi kèm nội dung có thể đã bị chỉnh sửa hoặc thao túng. Các dấu hiệu có thể xảy ra bao gồm cong vênh nơi các đường thẳng trên nền bây giờ xuất hiện gợn sóng, các bóng lạ, các cạnh lởm chởm hoặc màu da trông quá hoàn hảo. Cũng nên nhớ rằng, một hình ảnh có thể chính xác nhưng được sử dụng đơn giản trong bối cảnh gây hiểu nhầm. Chúng ta có thể sử dụng các công cụ như Google's Reverse Image Search của Google để kiểm tra xem hình ảnh có nguồn gốc từ đâu và hình ảnh đó có bị thay đổi hay không.

3. Phát hiện tự động tin giả

3.1. Khái niệm

Việc phát hiện tin giả một cách thủ công thường liên quan đến tất cả các kỹ thuật và quy trình mà một người có thể sử dụng để xác minh tin tức. Tuy nhiên, lượng dữ liệu trực tuyến được tạo ra hàng ngày là quá lớn. Hơn nữa,

thông tin lan truyền trực tuyến rất nhanh nên việc kiểm tra thủ công nhanh chóng trở nên không hiệu quả và thiếu thực tế. Việc kiểm tra thủ công gặp khó khăn lớn nhất khi mở rộng quy mô xác minh do khối lượng dữ liệu được tạo ra quá lớn và nhanh. Do đó, nhiệm vụ phát hiện tự động tin giả là một nhu cầu cấp bách và quan trọng.

Các nghiên cứu trước đây đã cho thấy sự khác biệt về khái niệm cũng như sự tương đồng giữa nhiều thuật ngữ liên quan đến "tin giả". Bên cạnh đó, các nghiên cứu cũng chỉ ra các phương pháp để xác minh tin giả. Tuy nhiên, để phát hiện tự động tin giả thì cần phải có các nghiên cứu sâu hơn. Các nghiên cứu hiện nay đang tiến thêm một bước nữa bằng cách xác định các đặc điểm hoặc chỉ số hoạt động cụ thể liên quan đến bản tin để trên cơ sở đó có thể mã hoá và đưa vào thuật toán học máy nhằm phân biệt một cách đáng tin cậy giữa các loại nội dung khác nhau được gắn với nhãn là "tin tức giả mạo".

Hệ thống phát hiện tự động tin giả sẽ giúp xác minh một tin tức là giả hay thật mà không cần sự can thiệp trực tiếp của con người. Có nhiều kỹ thuật và cách tiếp cận khác nhau được sử dụng trong nghiên cứu phát hiện tin giả. Các kỹ thuật và cách tiếp cận này phụ thuộc vào quan điểm và mục đích truy vết của người phát triển.

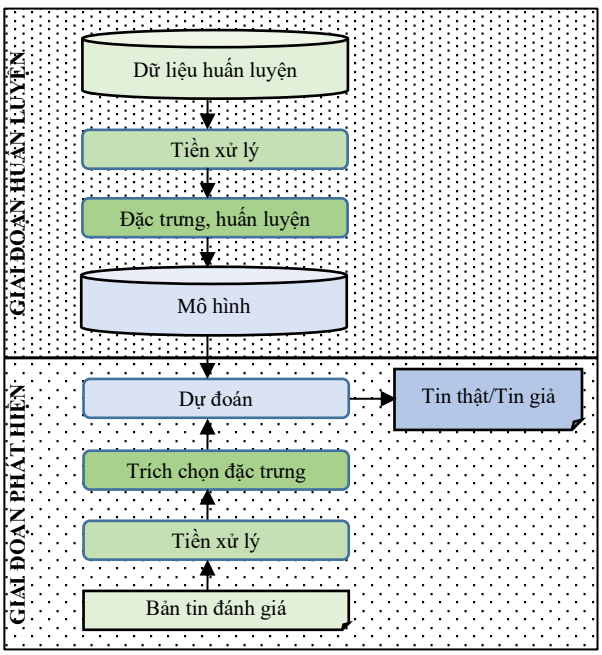
Trong bài báo này, nhóm tác giả chỉ giới thiệu hướng tiếp cận khá phổ biến hiện nay là dựa trên các kỹ thuật học máy (Machine Learning) với các phương pháp truyền thống (Naïve Bayes, Decision Tree, SVM, KNN) và dựa trên học sâu (Deep Learning). Các phương pháp này đều dựa trên phân tích nội dung để dự đoán tin giả.

3.2. Dựa trên các kỹ thuật học máy truyền thống

3.2.1. Phương pháp

Đa số các nghiên cứu theo hướng này đều sử dụng phương pháp học máy giám sát (Supervised Learning) hoặc học máy bán giám sát (Semi-Supervised Learning) để huấn luyện tạo mô hình nhằm mục đích phân loại tập dữ liệu và dự đoán.

Mô hình tổng quát của hướng tiếp cận này như Hình 2.



Hình 2. *Mô hình học máy để phát hiện tin giả*

Bước đầu tiên trong mô hình này là giai đoạn thu thập tập dữ liệu để xây dựng cơ sở dữ liệu huấn luyện. Trong cơ sở dữ liệu này bao gồm các bản tin đã được gán nhãn là tin giả hoặc tin thật. Trong trường hợp học máy giám sát, tất cả các dữ liệu dùng để huấn luyện đều phải được gán nhãn, trong trường hợp học bán giám sát thì bao gồm cả dữ liệu đã gán nhãn và chưa gán nhãn.

Giai đoạn tiền xử lý cho phép sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên để làm sạch dữ liệu, loại bỏ các thông tin không có ích và biểu diễn lại dữ liệu.

Giai đoạn trích chọn đặc trưng cho phép trích lọc những đặc trưng ngôn ngữ cần thiết phục vụ cho việc phân loại, nhận dạng nội dung. Trên cơ sở các đặc trưng đã trích xuất, thực hiện việc huấn luyện theo các thuật toán lựa chọn để xây dựng mô hình đặc trưng. Mô hình này sẽ được sử dụng cho giai đoạn dự đoán một bản tin là tin giả hay tin thật.

Giai đoạn dự đoán có chức năng đối sánh các đặc trưng của bản tin cần đánh giá với mô hình đặc trưng đã tạo ra trong giai đoạn huấn luyện để quyết định xem bản tin đó là tin giả hay tin thật.

Có nhiều thuật toán được sử dụng để huấn luyện và dự đoán trong học máy như [16]:

1) Naïve Bayes: Thuật toán này hoạt động dựa trên tiếp cận xác suất và định lý Bayes. Nói một cách đơn giản, Naïve Bayes giả định rằng, một thuộc tính trong danh mục này không liên quan gì đến các thuộc tính khác. Ví dụ, trái cây sẽ được phân loại là táo khi có màu đỏ, hình xoáy và đường kính gần 8 cm. Bất kể các thuộc tính này phụ thuộc vào nhau hay các thuộc tính khác. Thuật toán này thường được sử dụng để phân loại văn bản.

Xác suất Naïve Bayes:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$$

- P(c|x) là xác suất của c khi biết x. Trong đó, c là các lớp (nhãn) và x là tập các thuộc tính (đặc trưng).

- P(c) là xác suất của lớp c.

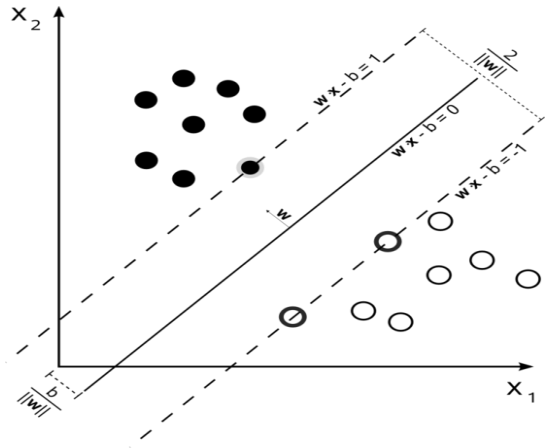
- P(x|c) là xác suất của x nếu biết c.

- P(x) là xác suất của x.

2) Decision Tree: Cây quyết định là một công cụ quan trọng hoạt động dựa trên cấu trúc giống như biểu đồ luồng được sử dụng chủ yếu cho các bài toán phân loại. Mỗi nút bên trong của cây quyết định chỉ định một điều kiện hoặc một "kiểm tra" trên một thuộc tính và việc phân nhánh được thực hiện trên cơ sở các điều kiện và kết quả kiểm tra. Cuối cùng, nút lá mang nhãn lớp thu được sau khi tính toán tất cả các thuộc tính. Khoảng cách từ gốc đến lá thể hiện quy luật phân loại. Chúng rất quan trọng trong việc tạo ra các biến và tính năng mới hữu ích cho việc khám phá dữ liệu và dự đoán biến mục tiêu khá hiệu quả. Đây là thuật toán phổ biến nên chi tiết về thuật toán tham khảo tại [16] cũng như các tài liệu khác.

3) Support Vector Machine (SVM): Đây là thuật toán hỗ trợ phân loại rất phổ biến và hiệu quả, có thể áp dụng trong học có giám sát hoặc bán giám sát. Mục đích của SVM là phân loại dữ liệu thành hai lớp khác nhau, trong

trường hợp này là lớp các tin giả và lớp các tin thật. Với một bộ các mẫu huấn luyện thuộc hai thể loại cho trước, thuật toán huấn luyện SVM xây dựng một mô hình SVM để phân loại các mẫu khác vào một trong hai thể loại đó. Thuật toán SVM chia hai lớp dữ liệu bằng một siêu mặt phẳng d–1 chiều khi số chiều của dữ liệu huấn luyện là d. Trong đó, w.x–b=0 là siêu mặt phẳng thể hiện sự phân tách dữ liệu.



Hình 3. *Mô hình phân lớp bằng SVM*

4) K-Nearest Neighbors (KNN): Một thuật toán đơn giản được sử dụng cho cả nhiệm vụ phân loại và hồi quy. KNN là một kĩ thuật học có giám sát (supervised learning) dùng để phân loại quan sát mới bằng cách tìm điểm tương đồng giữa quan sát mới này với dữ liệu sẵn có.

Ý tưởng của thuật toán KNN cho rằng, những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm dữ liệu x, y có k thuộc tính có thể dựa trên các khoảng cách:

- Euclidean: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$;

- Manhattan: $\sum_{i=1}^{k}|x_i - y_i|$;

- Minkowski: $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$.

3.2.2. Ưu điểm

Các phương pháp phát hiện tin giả dựa trên học máy để phân tích nội dung có một số ưu điểm nhất định như:

- Dễ dàng triển khai vì các giải thuật học máy đã được ứng dụng rất nhiều và đã có những cải tiến, hoàn thiện nhất định để đảm bảo chất lượng tốt và thời gian phân tích ngắn.

- Cho kết quả khá tốt trong trường hợp có một bộ dữ liệu chất lượng và được cập nhật thường xuyên. Việc phát hiện tin giả chỉ đơn thuần là thực hiện phân loại nhị phân vào một trong hai nhóm là tin giả hoặc tin thật.

- Đây là hướng tiếp cận phát hiện tin giả dựa trên nội dung nên có thể kết hợp với các hướng tiếp cận khác như phân tích về lan truyền tin, mức độ trích dẫn, phân tích hình ảnh... để nâng cao hơn độ chính xác của phát hiện tin giả.

3.2.3. Hạn chế

Tuy nhiên, phương pháp học máy có một số hạn chế cần phải tiếp tục nghiên cứu, khắc phục bao gồm:

- Phải liên tục cập nhật dữ liệu huấn luyện để điều chỉnh mô hình vì trong thực tế dữ liệu tin tức thay đổi hàng ngày, hàng giờ, thậm chí hàng giây.

- Phương pháp này chỉ mới dừng ở phân tích nội dung mà chưa tính đến các yếu tố khác như đặc điểm lan truyền tin, hình ảnh,...

- Phụ thuộc vào lĩnh vực thông tin như chính trị, dịch bệnh, quảng cáo,... nên cần phân loại dữ liệu trước để tăng tốc độ xử lý và độ chính xác.

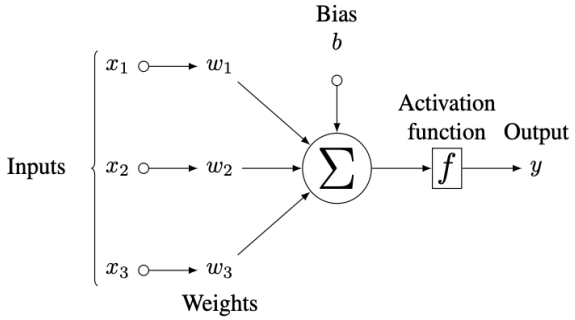
3.3. Dựa trên các kỹ thuật học sâu

3.3.1. Phương pháp

Các mạng nơ-ron chuyển tiếp sâu được gọi là mạng nơ-ron truyền thẳng hoặc các perceptron nhiều lớp là các mô hình cơ bản của học sâu.

Mục tiêu của mạng nơ-ron chuyển tiếp là làm gần đúng hàm f^*. Ví dụ, $y=f^*(x)$ ánh xạ đầu vào x (input) thành đầu ra y (output). Mạng nơ-ron chuyển tiếp (forward neural network) định nghĩa một ánh xạ $y=f(x; b)$ và tìm giá trị của các tham số b, dẫn đến giá trị xấp xỉ tốt nhất của hàm f.

Mô hình tổng quát của mạng nơ-ron được biểu diễn như Hình 4.



Hình 4. *Mô hình mạng nơ-ron*

Mô hình cơ bản của một tế bào nơ-ron được gọi là tế bào cảm thụ. Perceptron nhận tín hiệu đầu vào $x= (x_1, x_2, ..., x_{n+1})$ thông qua các lớp chuyển tiếp để tạo ra véc-tơ $w=(w_1, w_2, ..., w_{n+1})$. Đầu ra Perceptron được cho dưới dạng tích vô hướng của trọng số và véc-tơ, được biến đổi bởi hàm kích hoạt:

$$output = f(w.x) = f(\sum_{i=1}^{n+1} w_i x_i)$$

Dựa trên mô hình tổng quát này, người ta có thể đề xuất các thuật toán học sâu khác nhau hoạt động tương tự như các thuật toán học máy. Tuy nhiên, có một sự khác biệt chính đó là các thuật toán học sâu có các lớp diễn giải dữ liệu khác nhau. Mạng nơron nhân tạo đề cập đến mạng của các thuật toán như vậy (gọi chung là Perceptron) [17].

1) Mạng nơ-ron tích chập

Mạng nơ-ron tích chập (Convolutional neural networks - CNN) là mạng nơ-ron đặc biệt được sử dụng để xử lý dữ liệu. Những dữ liệu này được biểu diễn chính dưới dạng ma trận. Dữ liệu trong trường hợp phát hiện tin giả là tập hợp m văn bản, mỗi văn bản được xử lý và biểu diễn dưới dạng một véc-tơ n chiều. Như vậy, dữ liệu vào sẽ là một ma trận m×n.

Các mạng tích chập nằm trong các mạng nơron đơn giản sử dụng tích chập của nhiều dữ liệu số có thể có trong một trong các lớp của chúng.

Khái niệm tích chập trong toán học được định nghĩa là:

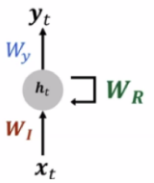
$$S(i,j) = (K * I)(i,j) = \sum_{i=1}^{m} \sum_{j=1}^{n} I(m,n)K(i-m, j-n)$$

Các lớp tích chập (convolutional layer) có các tham số K (Kernel) đã được học để tự điều chỉnh và lấy ra những thông tin chính xác nhất mà không cần chọn các đặc trưng.

2) Mạng nơ-ron hồi quy

Trong mô hình mạng nơ-ron thông thường, chúng ta coi input là các dữ liệu độc lập, không có mối liên hệ với nhau. Tuy nhiên, trong ngôn ngữ tự nhiên thì mối liên hệ giữa các từ và ngữ cảnh đóng một vai trò quan trọng, quyết định ý nghĩa của câu văn. Do đó việc áp dụng mô hình mạng nơ-ron thông thường vào các bài toán xử lý ngôn ngữ tự nhiên thường không đạt kết quả mong muốn.

Để khắc phục nhược điểm này, chúng ta sử dụng mô hình mạng nơ-ron hồi quy (Recurrent Neural Network - RNN). RNN coi dữ liệu đầu vào là một chuỗi liên tục và có thứ tự (Sequence), nối tiếp nhau theo thứ tự thời gian. Ví dụ như một đoạn văn bản có thể được coi là một chuỗi các từ (words) hoặc là một chuỗi các ký tự (character). Tại thời điểm t, với dữ liệu đầu vào x_t ta có kết quả output là y_t. Tuy nhiên, khác với mạng nơ-ron thường, y_t lại được sử dụng là input để tính kết quả output cho thời điểm (t+1). Điều này cho phép RNN có thể lưu trữ và truyền thông tin đến thời điểm tiếp theo. Mô hình hoạt động của RNN có thể được mô tả trong Hình 5.



$$h_t = g_h(W_1 x_t + W_R h_{t-1} + b_h)$$
$$y_t = g_y(W_y h_t + b_y)$$

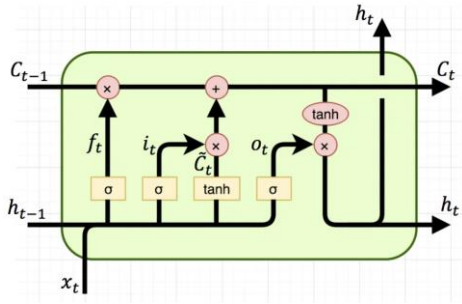
Hình 5. *Mô tả cách xử lý của mạng RNN*

Thông thường hàm kích hoạt g_h được sử dụng là *tanh* còn g_y có thể là hàm *sigmoid* hoặc *softmax* tùy thuộc vào từng bài toán cụ thể.

3) Mạng bộ nhớ ngắn-dài hạn LSTM

Về mặt lý thuyết thì RNN có thể xử lý và lưu trữ thông tin của một chuỗi dữ liệu với độ dài bất kỳ. Tuy nhiên, trong thực tế thì RNN chỉ tỏ ra hiệu quả với chuỗi dữ liệu có độ dài không quá lớn (short-term memory). Nguyên nhân của vấn đề này là do vấn đề suy giảm gradient (gradient được sử dụng để cập nhật giá trị của ma trận trọng số trong RNN và nó có giá trị nhỏ dần theo từng lớp khi thực hiện lan truyền). Khi gradient trở nên rất nhỏ (có giá trị gần bằng 0) thì giá trị của ma trận trọng số sẽ không được cập nhật thêm và do đó mạng Neuron sẽ dừng việc học tại lớp này. Đây cũng chính là lý do khiến cho RNN không thể lưu trữ thông tin của các bước thời gian trước đó trong một chuỗi dữ liệu có độ dài lớn.

LSTM (Long Short Term Memory) là một mạng cải tiến của RNN nhằm giải quyết vấn đề ghi nhớ lại giá trị các lớp trước đó. Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó, mỗi nút mạng đã có thể ghi nhớ được mà không cần bất kì can thiệp nào. Chi tiết về cách thức xử lý tại một nút mạng của LSTM được mô tả như Hình 6 [17].



Hình 6. *Mô tả một nút mạng trong LSTM*

Trong đó, f_t, i_t, o_t tương ứng với forget gate (cổng quên), input gate (cổng vào) và output gate (cổng ra).

- Cổng vào: Bước này sẽ quyết định xem thông tin nào cần bỏ đi từ trạng thái tế bào. Nó sẽ lấy đầu vào là h_{t-1} và x_t rồi đưa ra kết quả là một số trong khoảng [0, 1] cho mỗi số trong trạng thái tế bào C_{t-1}. Nếu là 1 nó sẽ lưu trữ thông tin lại cho sau này, còn 0 sẽ xoá toàn bộ thông tin.

Hàm f_t được tính như sau:

$$f_t = \sigma(W_f * x_t + W_f * h_{t-1} + b_f)$$

- Cổng vào: Bước này quyết định xem thông tin mới nào ta sẽ lưu vào trạng thái tế bào. Việc này gồm 2 phần: Đầu tiên là sử dụng một tầng *sigmoid* được gọi là "tầng cổng vào" để quyết định giá trị nào ta sẽ cập nhập; Tiếp theo là một tầng *tanh* tạo ra một véc-tơ cho giá trị mới $\tilde{C}_t$ nhằm thêm vào cho trạng thái. Sau đó, mạng sẽ kết hợp 2 giá trị đó lại để tạo ra một cập nhập cho trạng thái.

$$i_t = \sigma(W_i * x_t + W_i * h_{t-1} + b_i)$$
$$\tilde{C}_t = \tanh(W_c * x_t + W_c * h_{t-1} + b_c)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t)$$

- Cổng ra: Giá trị đầu ra sẽ dựa vào trạng thái tế bào, nhưng sẽ được tiếp tục sàng lọc. Đầu tiên, ta chạy một tầng *sigmoid* để quyết định phần nào của trạng thái tế bào ta muốn xuất ra. Sau đó, cung cấp trạng thái tế bảo qua một hàm *tanh* để có giá trị trong khoảng [-1,1] và nhân nó với đầu ra của cổng *sigmoid* để được giá trị đầu ra mong muốn.

$$o_t = \sigma(W_0 * x_t + W_o * h_{t-1} + b_0)$$
$$h_t = o_t * x_t + \tanh(C_t)$$

3.3.2. Ưu điểm

Học sâu là một dạng đặc biệt của học máy nên có tất cả ưu điểm của học máy như đã trình bày ở Mục 3.2.2. Tuy nhiên, khi ứng dụng các kỹ thuật học sâu để phát hiện tin giả, ta có thể áp dụng các kỹ thuật xử lý ở nhiều tầng tương ứng với các lớp mạng thay vì chỉ xử lý tuyến tính như các kỹ thuật học máy. Hơn nữa, các kỹ thuật học sâu cho phép xử lý khối lượng dữ liệu rất lớn và vì vậy rất phù hợp với phát hiện tin giả trong bối cảnh dữ liệu gia tăng nhanh chóng hiện nay.

3.3.3. Hạn chế

Với kết quả nghiên cứu đến lúc này, việc ứng dụng học sâu trong phát hiện tin giả vẫn gặp phải các vấn đề như đối với các phương pháp học máy đã trình bày ở Mục 3.2.3.

Hơn nữa, việc tự phát triển và cài đặt một hệ thống dựa trên học sâu là khá phức tạp và tốn kém. Đa số các thử nghiệm hiện nay trên các mạng học sâu được đều sử dụng

các bộ thư viện sẵn có của Google (Keras, TensorFlow), Facebook (Pytorch, Caffe), Microsoft (CNTK, Gluon),...

Cuối cùng, vấn đề làm thế nào để thiết kế một mạng học sâu xử lý tích hợp các thông tin dựa trên nội dung, phương thức lan truyền, văn phong,... là vẫn còn ở phía trước.

4. Thực nghiệm

4.1. Ngôn ngữ lập trình và môi trường thử nghiệm

Để phát triển các mô-đun chương trình, nhóm tác giả sử dụng ngôn ngữ lập trình Python. Python là ngôn ngữ được sử dụng phổ biến nhất trong Deep Learning và thư viện Deep Learning được chọn để sử dụng là Keras.

Môi trường đươc chọn để thử nghiệm là Google Colab (Google Colaboratory) vì đây là một dịch vụ miễn phí của Google nhằm hỗ trợ nghiên cứu và học tập về trí tuệ nhân tạo, có GPU để chạy các chương trình Python và hỗ trợ Deep Learning.

Đặc biệt, trên môi trường Colaboratory có cài sẵn các thư viện Deep Learning phổ biến như PyTorch, TensorFlow, Keras,... Ngoài ra, ta cũng có thể cài thêm các thư viện khác để chạy nếu cần. Nhóm tác giả thực hiện liên kết Google Colaboratory với Google Drive để lưu trữ và truy xuất dữ liệu nên rất tiện để sử dụng.

4.2. Xây dựng phần mềm phân loại

Để xây dựng phần mềm phân loại văn bản tiếng Việt, nhóm tác giả thực hiện qua các bước sau:

- Chuẩn bị dữ liệu (bao gồm dữ liệu huấn luyện và dữ liệu để thử nghiệm việc phân loại);

- Tiền sử lý dữ liệu;

- Xây dựng mô hình (thông qua việc huấn luyện để tạo mô hình và tinh chỉnh);

- Xây dựng phần mềm phân loại văn bản.

4.3. Chuẩn bị dữ liệu

Để kiểm chứng các phương pháp trình bày ở trên, nhóm tác giả đã tiến hành thử nghiệm với bộ dữ liệu gồm các tin chính trị (kênh chính thống lấy từ báo Nhân dân, Thông tấn xã và một số báo khác; kênh tin giả lấy từ một số Blog, Facebook).

Bảng 1. *Dữ liệu huấn luyện và thử nghiệm*

TT	Thể loại	Dữ liệu huấn luyện		Dữ liệu thử nghiệm	
		Tập tin	Kích thước (MB)	Tập tin	Kích thước (MB)
1	Chính trị (tin thật)	5.220	22,4	7.569	32,9
2	Chính trị (tin giả)	3.160	16,8	2.037	13,9
3	Covid (tin thật)	1.821	8,1	2.097	10,0
4	Covid (tin giả)	2.553	10,8	5.277	24,7

4.4. Kết quả thử nghiệm

Kết quả thử nghiệm thu được đối với từng phương pháp như Bảng 2.

Bảng 2. *So sánh kết quả thử nghiệm*

Lĩnh vực	SVM			CNN			RNN		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Chính trị	0,68	0,67	0,67	0,69	0,68	0,68	0,78	0,77	0,77
Covid-19	0,73	0,51	0,60	0,76	0,55	0,64	0,83	0,65	0,73

5. Một số thách thức

Qua nghiên cứu các tài liệu đã công bố và một số công cụ thử nghiệm đã có, nhóm tác giả thấy nổi lên một số thách thức và đó cũng là những hướng nghiên cứu tiềm năng trong tương lai đối với phát hiện tin giả, đặc biệt là đối với tiếng Việt.

- *Vấn đề về sưu tập dữ liệu*: Muốn có được một hệ thống phát hiện tin giả dựa trên nội dung thì trước hết phải có bộ dữ liệu huấn luyện đủ lớn và được cập nhật kịp thời. Tuy nhiên, trong thực tế thì khối lượng thông tin phát sinh trên mạng là cực kỳ lớn và thay đổi theo thời gian thực. Vì vậy, việc thu thập, phân loại và gán nhãn cho những dữ liệu này là một thách thức lớn.

- *Vấn đề cập nhật lại mô hình đặc trưng*: Nếu cập nhật thường xuyên dữ liệu huấn luyện thì sẽ kéo theo vấn đề phải cập nhật lại mô hình đặc trưng. Trong trường hợp này việc tính toán và cập nhật lại mô hình đặc trưng thế nào cho nhanh và bảo đảm hiệu quả là một vấn đề cần nghiên cứu.

- *Vấn đề thích ứng nội dung trên các tin giả.* Các đối tượng phát tán tin giả ngày càng tinh vi và luôn biết cách viết các tin giả giống như thật. Họ luôn tìm cách cải thiện kỹ năng viết tin giả để qua mặt các hệ thống phát hiện tin giả và đây là một thách thức không nhỏ cần phải vượt qua.

- *Vấn đề tích hợp các yếu tố để xác định tin giả*: Như đã thảo luận ở các phần trước, phân tích nội dung chỉ là một trong yếu tố để có thể xác định tin giả. Các yếu tố khác như nguồn phát tán tin, cách thức lan truyền tin, nội dung các bình luận liên quan đến bản tin, văn phong sử dụng trên bản tin,... đều có thể là các manh mối quan trọng để xác định tin giả. Làm thể nào để kết hợp được các yếu tố này khi phát hiện tin giả cũng là một vấn đề quan trọng cần phải nghiên cứu.

6. Kết luận

Trong những năm gần đây, sự phát triển của mạng Internet và đặc biệt là mạng xã hội trực tuyến đã tạo điều kiện thuận lợi hơn rất nhiều để mọi người giao tiếp với nhau qua mạng. Người dùng dễ dàng chia sẻ thông tin, kết nối với những người khác và cập nhật thông tin về các sự kiện diễn ra hàng ngày. Tuy nhiên, cùng với những tiện ích to lớn thì song hành với nó là sự xuất hiện nhiều vấn nạn mới, đặc biệt là vấn nạn về tin giả. Một lượng lớn tin tức giả trên mạng tiềm ẩn nguy cơ gây ra nhiều vấn đề nghiêm trọng trong xã hội. Việc xử lý vấn đề tin giả đã và đang thu hút sự chú ý của ngành công nghiệp và giới học thuật nhằm tìm hiểu về nguồn gốc, sự phân bố, tác hại và ngăn chặn chúng.

Bài báo mang tính chất nghiên cứu tổng quan, nhóm tác giả đã trình bày các vấn đề liên quan đến khái niệm, phân

loại, cách xác định thủ công và xác định tin giả tự động. Đặc biệt, nhóm tác giả đã trình bày hai kỹ thuật ứng dụng rộng rãi hiện nay đó là kỹ thuật học máy dựa trên các phương pháp truyền thống và học sâu. Hai kỹ thuật này đều dựa trên học máy để phân tích nội dung và bước đầu đã mang lại những kết quả tích cực.

Tuy nhiên, để xác định chính xác và nhanh một tin tức có phải là giả hay không vẫn còn rất nhiều thách thức. Trong thời gian đến, nhóm tác giả sẽ nghiên cứu đề xuất một hệ thống xác định tin giả cho tiếng Việt, kết hợp phân tích nội dung với các yếu tố khác như nguồn gốc phát tán, phương thức lan truyền, văn phong,...

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Bộ Giáo dục và Đào tạo thông qua đề tài có mã số B2022-DNA-17.

TÀI LIỆU THAM KHẢO

[1] Dat Quoc Nguyen and Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese", *Proceedings EMNLP 2020 (The 2020 Conference on Empirical Methods in Natural Language Processing)*, 2020, pages 1037–1042.

[2] Duc-Trong Le et al., "ReINTEL: A multimodal data challenge for responsible information identification on social network sites", *7th Annual Workshop on Vietnamese Language and Speech Processing - VLSP 2020*, 2020, https://arxiv.org/pdf/2012.08895.pdf.

[3] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, Volume 3, No 2, 2017.

[4] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online", *Science Journals*, Volume 359, Issue 6380, 2018, p.p. 1146–1151.

[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective", *ACM SIGKDD, Explorations Newsletter*, Volume 19, No 1, 2017, p.p. 22–36.

[6] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences", *IJCAI 2017, Workshop on Explainable Artificial Intelligence (XAI)*, 2017, arXiv preprint arXiv:1712.00547.

[7] S.A. McCornack, K. Morrison, J.E. Paik, A.M. Wisner, and X. Zhu, "Information manipulation theory 2: A propositional theory of deceptive discourse production", *Journal of Language and Social Psychology*, Volume 33, No 4, 2014, p.p. 348–377.

[8] M. Zuckerman, B.M. DePaulo, and R. Rosenthal, "Verbal and Nonverbal Communication of Deception", *In Advances in experimental social psychology, Elsevier*, Volume 14, 1981, p.p. 1–59.

[9] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-Aware Multi-modal Fake News Detection", *In Advances in Knowledge Discovery and Data Mining, Springer International Publishing*, 2020, p.p. 354–367.

[10] X. Zhou and R. Zafarani, "Network-based Fake News Detection: A Pattern-driven Approach", *arXiv e-prints*, 2019, arXiv:1906.04210.

[11] D. Kollias, S.P. Zafeiriou, "Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the OMG in-the-wild dataset*", IEEE Transactions on Affective Computing*, 2020.

[12] M.K. Elhadad, K.F. Li, F. Gebali, "A novel approach for selecting hybrid features from online news textual metadata for fake news detection", *International conference on p2p, parallel, grid, cloud and internet computing, Springer*, 2019, p.p. 914-925.

[13] N. Higdon, "*The anatomy of fake news: A critical news education*", Oakland, CA: University of California Press, 2020.

[14] C. Wardle, H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making", *Report of DGI (Directorate General Human Rights and Rule of Law)*, Published by the Council of Europe, 2017.

[15] D. Fallis, "What Is Disinformation?", *Library Trends, Johns Hopkins University Press*, Volume 63, Number 3, 2015, pp. 401-426.

[16] Z. Khanam, B.N. Alwasel, H. Sirafi and M. RashidFake, "News Detection Using Machine Learning Approaches", *IOP Conference Series: Materials Science and Engineering*, Published by IOPscience, 2021, DOI: 10.1088/1757-899X/1099/1/012040.

[17] V.M. Kresnakova, M. Sarnovsky, P. Butka, "Deep learning methods for Fake News detection", *Proceedings of 19th International Symposium on Computational Intelligence and Informatics*, IEEE, 2020, DOI: 10.1109/CINTI-MACRo49179.2019.9105317.