Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
Năm thứ 33, Số 5 (2022), 26–46

www.jabes.ueh.edu.vn

JABES
Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á

Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á

http://www.emeraldgrouppublishing.com/services/publishing/jabes/

Nghiên cứu giải pháp phát hiện tin giả trên mạng xã hội bằng ngôn ngữ tiếng Việt

TRƯƠNG QUỐC ĐỊNH [a], PHAN THỊ THÚY KIỀU [b,*]

[a] *Trường Đại học Cần Thơ*

[b] *Phân hiệu Trường Đại học Kinh tế TP. Hồ Chí Minh tại Vĩnh Long*

THÔNG TIN

Ngày nhận: 03/02/2021
Ngày nhận lại: 19/10/2021
Duyệt đăng: 16/02/2022

Mã phân loại JEL:
C61; C63; C67

Từ khóa:
Tin giả;
Mạng xã hội;
Phát hiện tin giả;
Tóm tắt;
Độ tương đồng cosine.

Keywords:
Fake news;
Social media;
Fake new detection;
Summary;
Cosine similarity.

TÓM TẮT

Trong nghiên cứu này, nhóm tác giả sử dụng giải pháp tìm kiếm thông tin để tìm kiếm các tin thật có nội dung tương tự với nội dung của tin cần kiểm tra, sau đó dùng độ đo cosine để đo đạt và đánh giá bản tin kiểm tra có phải là tin giả hay không. Bài nghiên cứu sử dụng hai bộ dữ liệu cho mục tiêu xác định giá trị các tham số của mô hình đề xuất cũng như thực nghiệm độ chính xác của mô hình. Bộ dữ liệu tin thật được thu thập từ các trang tin chính thống của Việt Nam là tập hợp tin thật. Tập dữ liệu kiểm thử được thu thập từ các bài đăng trên mạng xã hội vừa có tin thật và tin giả dùng cho mục đích kiểm tra độ chính xác của mô hình đề xuất. Kết quả thực nghiệm trên hai bộ dữ liệu cho thấy rằng mô hình đề xuất của bài nghiên cứu có thể phát hiện tin giả.

Abstract

In this research, the authors use the information search solution to find real with similar content to the content of the newsletter to be tested, then use cosine measurement to measure and evaluate the news of the test is fake or not. This research uses two datasets for the purpose of determining the value of the parameters of the proposed model as well as testing the accuracy of the model. The real data set collected from the official Vietnamese news sites is the real one. Test datasets are collected from social media posts that have both real and fake news for the purpose of testing the accuracy of the proposed model. Experimental results on the two datasets show that our proposed model can detect fake news.

[*] Tác giả liên hệ.
Email: tqdinh@cit.ctu.edu.vn (Trương Quốc Định), kieuptt@ueh.edu.vn (Phan Thị Thúy Kiều).
Trích dẫn bài viết: Trương Quốc Định, & Phan Thị Thúy Kiều. (2022). Nghiên cứu giải pháp phát hiện tin giả trên mạng xã hội bằng ngôn ngữ tiếng Việt. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á, 33*(5), 26–46.

1. Giới thiệu

Cùng với sự phát triển của khoa học công nghệ, mà đặc biệt là lĩnh vực Công nghệ Thông tin và Truyền thông, đã hình thành phương thức trao đổi thông tin mới, phương thức trao đổi thông tin trực tuyến thông qua mạng xã hội. Hiện nay, mạng xã hội được nhiều người sử dụng rộng rãi, mạng xã hội có ưu điểm là lan truyền thông tin một cách nhanh chóng và tiếp cận đến nhiều người hơn các hình thức truyền thống khác. Mặt khác, mạng xã hội cũng giúp người dùng thể chia sẻ, đưa ra ý kiến bình luận và thực hiện thảo luận với bạn bè hoặc người dùng khác một cách dễ dàng hơn. Tuy nhiên, trong vài năm trở lại đây, đặc biệt là kể từ cuộc bầu cử Tổng thống Mỹ năm 2016, tin giả trên mạng xã hội xuất hiện ngày càng nhiều và phát triển với tốc độ nhanh chưa từng có. Tin giả thường là các tin có nội dung "giật gân", "ly kỳ" được sử dụng cho nhiều mục đích khác nhau như: Thu hút sự chú ý của người dùng để tăng ảnh hưởng của người đăng trên mạng xã hội, tăng lượt truy cập để tăng doanh thu… Tuy nhiên, việc phát tán và lan truyền tin giả, tin có nội dung không đúng sự thật thường tổn hại đến uy tín của cá nhân và tổ chức, đôi khi gây bất ổn về trật tự xã hội cũng như lợi ích chính trị của quốc gia, dân tộc (Thanh Hà, 2018).

Tin giả cũng gây ảnh hưởng không nhỏ đối với cá nhân, tổ chức và doanh nghiệp. Một ví dụ điển hình như: Trên mạng xã hội Facebook tại Việt Nam xuất hiện trang fanpage mạo danh "Báo Công an", trong đó đăng clip về vụ việc một Việt kiều về nước, tranh cãi với nhân viên sân bay vì cho rằng hành lý của họ bị cắt khóa, lấy đồ tại sân bay Việt Nam, nhưng thực chất không phải như vậy. Mạng xã hội cũng xuất hiện trang "Cục Hải quan bán xe thanh lý", trong đó không chỉ sử dụng logo của Hải quan làm ảnh đại diện mà còn sử dụng ảnh chụp tập thể cán bộ, nhân viên của Báo Hải quan làm ảnh nền. Đối tượng xấu còn đăng hình ảnh về dàn siêu xe gắn biển xanh (thực chất là đồ chơi bằng nhựa) lên các trang mạng gây hiểu lầm, tạo phản ứng xấu trong dư luận về xe công vụ (Thanh Hà, 2018).

Ngoài ra, nhiều tổ chức là các doanh nghiệp, nhãn hàng lớn cũng điêu đứng vì tin tức giả trên Facebook và Youtube. Chẳng hạn như: Video gây tổn hại nghiêm trọng đến uy tín và doanh số bán hàng của Heineken Việt Nam được lan truyền trên Facebook và Youtube từ ngày 04/5/2017, hay các fanpage mạo danh các hãng ô tô lớn như: Toyota, Honda, Kia Morning đều lan truyền các tin tức kiểu tặng xe cho những người may mắn nhân dịp các sự kiện quan trọng của công ty (Bông Mai, 2017).

Gần đây, lợi dụng diễn biến phức tạp của dịch COVID-19, nhiều cá nhân đã đăng những thông tin sai sự thật, giật gân, câu like trên các trang mạng xã hội Facebook gây khó khăn cho công tác phòng, chống dịch. Một số tin giả liên quan đến dịch COVID-19 như: Theo Vietnamnet[1], bài viết trên Facebook có nội dung "Đêm nay nhà nước mình phun thuốc ngừa dịch cúm corona trên bầu trời toàn quốc nên mọi người hãy chia sẻ cho nhau. Không nên ra đường trong khung giờ từ 4–7g30 sáng ngày 01/02/2020, nếu có việc phải ra đường thì nên đeo khẩu trang". Theo Tuổi Trẻ[2], bài viết có nội dung "TP.HCM bị phong tỏa từ ngày 28/3..." và còn rất nhiều tin giả khác đang tràn lan trên các trang mạng xã hội.

Hiện nay, các nước trên thế giới đều đã đưa ra luật về chống tin giả như: Singapore, Indonesia, Malaysia… Tại Việt Nam, căn cứ Khoản a Điều 101 Nghị định 15/2020/NĐ-CP của Chính phủ về xử phạt hành chính trong lĩnh vực bưu chính viễn thông, tần số vô tuyến điện (thay thế Nghị định

[1] Truy cập từ https://vietnamnet.vn/trieu-tap-co-gai-thanh-hoa-tung-tin-phun-thuoc-ngua-virus-corona-len-troi-613028.html
[2] Truy cập từ https://tuoitre.vn/bi-phat-10-trieu-dong-vi-tung-tin-gia-phong-toa-tp-hcm-14-ngay-20200329202822702.htm

174/2013) có hiệu lực từ tháng 4 năm 2020, người nào cung cấp, chia sẻ thông tin giả mạo, thông tin sai sự thật, xuyên tạc, vu khống, xúc phạm uy tín của cơ quan, tổ chức, danh dự, nhân phẩm của cá nhân thì phạt tiền từ 10 triệu đồng đến 20 triệu đồng. Ngoài ra, người đó phải gỡ bỏ thông tin sai sự thật, gây nhầm lẫn hoặc thông tin vi phạm pháp luật.

Tuy nhiên, theo Bộ trưởng Bộ Thông tin Truyền thông thừa nhận, việc quản lý, phát hiện và chống phát tán tin giả hiện nay vẫn còn nhiều hạn chế và bất cập. Ngoài hạn chế về kỹ thuật, Bộ trưởng cho biết, các đối tượng phát tán thông tin thường xuyên tận dụng những thay đổi, bước phát triển mới về công nghệ để cải tiến hình thức phát tán thông tin để thông tin được lan truyền với tốc độ nhanh và khó kiểm soát (Bông Mai, 2017).

Các phần tiếp theo của bài nghiên cứu sẽ làm sáng tỏ vấn đề này theo bố cục sau: Phần 2 trình bày cơ sở lý thuyết về tin giả; Phương pháp nghiên cứu được trình bày trong phần 3; Phần 4 sẽ phân tích và thảo luận kết quả; và cuối cùng phần 5 là trình bày kết luận.

2. Cơ sở lý thuyết

2.1. Định nghĩa tin giả

Tin giả (Fake News) đã tồn tại trong thời gian dài, tuy nhiên, chưa có định nghĩa chính xác về tin giả. Trong nghiên cứu này, nhóm tác giả giới thiệu một số định nghĩa về tin giả như sau:

Theo Allcott và Gentzkow (2017) định nghĩa "fake news as news stories that have no factual basis but are presented as facts", điều này có nghĩa là "tin giả là tin mà không có cơ sở thực tế nhưng được trình bài là tin thật".

Theo từ điển Collins[3], định nghĩa tin tức giả là "false and sometimes sensationalist information presented as fact and published and spread on the internet", có nghĩa là "những thông tin sai, thường là giật gân, được phát tán dưới vỏ bọc tin tức".

Theo Wikipedia[4], tin giả (Fake News) còn được gọi là tin rác hoặc tin tức giả mạo, là một loại hình báo chí hoặc tuyên truyền, bao gồm: Các thông tin cố ý hoặc trò lừa bịp lan truyền qua phương tiện truyền thông, tin tức truyền thống (in và phát sóng), hoặc phương tiện truyền thông xã hội trực tuyến.

Theo Shu và cộng sự (2017), tin giả được định nghĩa như sau "fake news is a news article that is intentionally and verifiably false" có nghĩa là "Tin giả là bài viết tin tức mà cố ý và được xác minh là sai". Trong nghiên cứu này, nhóm tác giả thực hiện nghiên cứu theo định nghĩa này nhằm phát hiện các tin giả trên mạng xã hội.

2.2. Một số đặc điểm nhận biết tin giả trên mạng xã hội

Có 3 đặc điểm cơ bản để phát hiện tin giả trên mạng xã hội là: User[5], Posts[6], và Network[7] (Shu và cộng sự, 2017).

[3] Truy cập từ https://www.collinsdictionary.com/us/submission/18357/fake+news
[4] Wikipedia: Bách khoa toàn thư mở. Truy cập từ https://vi.wikipedia.org/wiki/Tin_gi%E1%BA%A3
[5] User: Tài khoản
[6] Post: Các bài đăng
[7] Network: Mạng

User-based: Tin giả có thể được tạo và phát tán từ những tài khoản độc hại như: Bots, Cyborg hoặc Trolls[8]. Các tính năng dựa trên người dùng thể hiện những người dùng đó có tương tác với tin tức trên phương tiện truyền thông xã hội. Các đặc điểm của mạng xã hội có thể được chia thành các cấp độ khác nhau: Cấp độ cá nhân và cấp nhóm. Cấp độ cá nhân bao gồm các thông tin liên quan như: Tuổi, số người theo dõi/người theo dõi, số lượng bài đăng… Ở cấp độ nhóm, người dùng sẽ biết được các thông tin có liên quan đến tin tức của bài viết mà người dùng đăng trong nhóm (Ahmed, 2017).

Post-based: Mọi người bày tỏ ý kiến hoặc cảm xúc của họ thông qua các bài đăng trên mạng xã hội như: Ý kiến phản hồi, phản ứng giật gân… Vì vậy, việc trích xuất các đặc trưng Post-based giúp tìm thấy các tin tức giả thông qua các bài đăng công khai. Đặc trưng Post-based dựa trên xác thực thông tin người dùng để suy ra tính xác thực từ nhiều khía cạnh liên quan đến bài đăng trên mạng xã hội. Chúng ta có thể trích xuất các thông tin Post-based để phát hiện tin giả trên mạng xã hội. Các tính năng này được chia thành ba cấp độ: (1) Post Level, (2) Group Level, và (3) Temporal Level (Ahmed, 2017).

(1) Post Level: Thường là các bài đăng trên mạng xã hội. Mỗi bài đăng thường có các đặc điểm như: Quan điểm, chủ đề, độ tin cậy.

(2) Group Level: Tất cả các bài đăng có liên quan, cụ thể là bài viết sử dụng "Wisdom of Crowds" có nghĩa là "Trí khôn của đám đông". Ví dụ, trung bình các điểm tin cậy được sử dụng để đánh giá độ tin cậy của tin tức (Jin và cộng sự, 2016).

(3) Temporal Level: Nhằm xác định thời điểm đăng bài trên mạng xã hội, một số phương pháp được sử dụng như: Mạng nơ-ron hồi quy (Recurrent Neural Network – RNN) được sử dụng để thu hút các bài đăng thay đổi theo thời gian (Ma và cộng sự, 2016; Ruchansky và cộng sự, 2017). Dựa trên hình dạng của chuỗi thời gian này cho các số liệu khác nhau của các bài đăng có liên quan (ví dụ: Số lượng bài đăng), các tính năng toán học có thể được tính toán, chẳng hạn như: Các tham số theo thời gian (SpikeM) (Kwon và cộng sự, 2013).

- Network-based: Người dùng sử dụng mạng xã hội kết nối các thành viên có cùng sở thích, chủ đề và có mối quan hệ với nhau. Network-based trích xuất cấu trúc đặc biệt từ những người dùng đăng các bài viết công khai trên mạng xã hội. Network-based được xây dựng theo các kiểu khác nhau (Ahmed, 2017).

- Stance Network: Được xây dựng bởi các nút hiển thị cho tất cả các bài đăng có liên quan đến tin tức và các cạnh thể hiện trọng lượng của sự tương đồng Stance Network.

- Co-occurrence Network: Được xây dựng dựa trên tương tác của người dùng bằng cách đếm xem những người dùng đó có các bài đăng liên quan đến cùng một bài viết hay không.

- Friendship Network: Cho biết người dùng theo hay không theo bài viết có liên quan.

Dựa vào các đặc trưng phát hiện tin giả trên mạng xã hội đã được trình bày ở trên. Trong nghiên cứu này, nhóm tác giả sử dụng các đặc trưng của Post-based để xác định các tin cần kiểm chứng.

[8] Bots, Cyborg và Trolls: Tên các phần mềm gây nguy hiểm cho các trang mạng và máy tính. Truy cập từ https://eds.s.ebscohost.com/eds/pdfviewer/pdfviewer?vid=1&sid=160960ba-e66f-4e9b-935e-5ddf4c08e3c3%40redis

2.3. Nghiên cứu liên quan

Hiện nay, vấn đề về phát hiện tin giả thu hút sự quan tâm, nghiên cứu của nhiều học giả trong và ngoài nước.

Phương pháp được đề xuất bởi Horne và Adali (2017), sử dụng tiêu đề để phân biệt giữa bài viết giả và thật. Theo nghiên cứu của Horne và Adali (2017), tiêu đề tin tức giả có ít từ vựng và danh từ hơn, trong khi đó, nội dung có nhiều danh từ và động từ. Họ trích lọc từ các tính năng khác nhau được nhóm lại thành ba loại như sau: (1) Tính toán độ phức tạp và khả năng đọc của văn bản; (2) Các tính năng về tâm lý và đo lường quá trình nhận thức và mối quan tâm cá nhân nằm dưới các tác phẩm, chẳng hạn như: Số từ cảm xúc của từ ngữ; và (3) Các tính năng phong cách phản ánh phong cách của người viết và cú pháp của văn bản, chẳng hạn như: Số động từ và số danh từ.

Các tính năng được đề cập ở trên đã được sử dụng để xây dựng một mô hình phân loại máy học vectơ hỗ trợ (Support Vector Machine – SVM). Horne và Adali (2017) đã sử dụng tập dữ liệu Buzzfeed[9] và các trang web tin tức khác cùng dữ liệu châm biếm của Burfoot và Baldwin (2009) để kiểm tra mô hình của Horne và Adali (2017) . Theo phương pháp của Horne và Adali (2017), so sánh với các bài báo châm biếm (bài báo hài hước) có độ chính xác là 91%. Tuy nhiên, độ chính xác giảm xuống còn 71% khi thực hiện dự đoán tin tức giả.

Nghiên cứu tiếp theo là phân loại tin đồn (Rumor Classification). Một tin đồn có thể sử dụng định nghĩa như "a piece of circulating information whose veracity status is yet to be verified at the time of spreading" có nghĩa là "một phần thông tin được lưu hành có trạng thái là chưa được xác minh tại thời điểm lan truyền" (Zubiaga và cộng sự, 2018). Chức năng của tin đồn được hiểu là một tình huống mơ hồ và giá trị thực có thể là đúng, sai hoặc chưa được xác minh. Trước đây, tin đồn tập trung vào bốn nhiệm vụ: (1) Phát hiện tin đồn, (2) theo dõi tin đồn, (3) phân loại lập trường, và (4) phân loại tính xác thực. Cụ thể, phát hiện tin đồn nhằm phân loại một mẩu tin là tin đồn hoặc không là tin đồn (Sampson và cộng sự, 2016; Wu và cộng sự, 2017). Nhiệm vụ liên quan đến phát hiện tin giả là phân loại tính xác thực tin đồn. Phân loại tính xác thực của tin đồn phụ thuộc rất nhiều vào các nhiệm vụ khác, lập trường hoặc ý kiến được trích xuất từ các bài viết có liên quan. Những bài đăng này là những tín hiệu quan trọng để xác định tính xác thực của tin đồn. Khác với tin đồn, có thể bao gồm tin đồn dài hạn, chẳng hạn như: Giả thuyết âm mưu, cũng như những tin đồn ngắn, hay các tin giả có đề cập đến các thông tin có liên quan, đặc biệt là các sự kiện tin tức công khai có thể được xác minh là sai.

Tiếp theo là nghiên cứu về khám phá sự thật (Truth Discovery) là vấn đề phát hiện sự thật từ nhiều nguồn xung đột (Li và cộng sự, 2016). Các phương pháp khám phá sự thật không trực tiếp khám phá các yêu cầu thực tế, mà dựa vào một tập hợp các nguồn mâu thuẫn ghi lại các thuộc tính của các đối tượng để xác định giá trị thật. Mục đích khám phá sự thật để xác định nguồn tin cậy và giá trị thực của đối tượng. Vấn đề phát hiện tin giả được hưởng lợi từ các khía cạnh khác nhau của phương pháp khám phá sự thật trong các tình huống khác nhau.

- *Đầu tiên,* độ tin cậy của các tin tức khác nhau có thể được mô hình hóa để suy ra tính trung thực của tin tức được báo cáo.

- *Thứ hai,* các bài đăng trên phương tiện truyền thông xã hội có liên quan cũng có thể được mô hình hóa như các nguồn phản hồi xã hội để xác định tốt hơn yêu cầu của tính trung thực (Mukherjee

[9] Buzzfeed: Dữ liệu phân tích tin giả và thật từ 2016 US Elections.

& Weikum, 2015; Weikum, 2017). Tuy nhiên, một số các vấn đề khác phải được xem xét để áp dụng khám phá sự thật để phát hiện tin tức giả trong các ngữ cảnh mạng xã hội. Đầu tiên, hầu hết các phương pháp khám phá sự thật hiện tại tập trung vào việc xử lý các đầu vào có cấu trúc dưới dạng Chủ ngữ - Vị ngữ - Tân ngữ (Subject-Predicate-Object – SPO), trong khi dữ liệu mạng xã hội có cấu trúc cao. Thứ hai, phương pháp khám phá sự thật không thể được áp dụng tốt khi một bài viết giả mới chỉ được tung ra và xuất bản bởi một vài cơ quan báo chí, vì tại thời điểm đó không có đủ các bài trên mạng xã hội có liên quan để làm nguồn bổ sung.

Một đề xuất của nhóm tác giả Aldwairi và Alwahedi (2018) về các nội dung tin giả gọi là Clickbait. Clickbait được thiết kế để tạo sự chú ý của người dùng khi nhấp vào các liên kết này sẽ chuyển hướng đến trang có tin giả. Mục đích của công việc nghiên cứu là đưa ra một giải pháp có thể được người dùng sử dụng để phát hiện và lọc ra các trang web có chứa thông tin sai lệch và gây hiểu lầm dựa trên Clickbait bằng việc ứng dụng máy học và các cách thức phân loại để phát hiện tin giả (WEKA Classifiers), bao gồm: BayesNet, Logistic, Random Tree và NaiveBayaes (mạng Bayes, hồi quy logistic, cây ngẫu nhiên và mô hình Naive Bayes) để phát hiện tin giả. Tóm lại, Aldwairi và Alwahedi (2018) sử dụng các tính năng đơn giản và được lựa chọn cẩn thận của tiêu đề và bài đăng để xác định chính xác các bài đăng giả mạo. Kết quả thử nghiệm cho thấy độ chính xác đến 99,4% bằng cách sử dụng lớp logistic. Phương pháp này thực hiện xác định xác suất các từ (Words) thuộc vào tin giả, từ đó xác định khả năng tin là thật hay giả.

Tại Việt Nam, một nghiên cứu về việc xác định tài khoản Spam trên mạng Facebook (Vương Thị Hồng và cộng sự, 2016). Bằng việc thu thập các bình luận và hành vi của người dùng trên mạng xã hội và sử dụng mô hình Entropy cực đại (Maximum Entropy) để phân lớp các bình luận và xác định xem tài khoản có spam hay không. Kết quả nghiên cứu cho thấy độ chính xác là 90%. Tuy nhiên, việc xác định các tin spam phụ thuộc vào các ràng buộc của Entropy.

Các nghiên cứu trên vẫn chưa xác định cụ thể tin giả, một số phương pháp sử dụng phương pháp máy học để xác định tin giả, tin sai sự thật. Hiện tại chưa có giải pháp áp dụng kiểm tra tin giả cho bộ dữ liệu tiếng Việt. Trong nghiên cứu này, nhóm tác giả không dựa vào máy học mà dựa vào các nguồn tài nguyên online được xác định là tin thật, từ đó, các tin không tương đồng với tin trong nguồn tài nguyên này thì được xem là tin giả hoặc không thật theo định nghĩa của Shu và cộng sự (2017).

2.4. Tóm tắt văn bản tự động

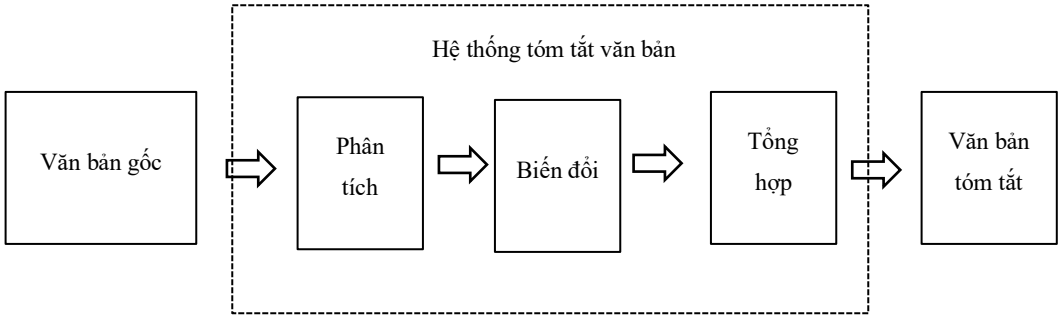
Tóm tắt văn bản đã trở thành một công cụ quan trọng và hữu ích để hỗ trợ và trích chọn thông tin văn bản trong thời đại thông tin phát triển nhanh chóng như ngày nay. Tóm tắt văn bản thủ công (được thực hiện bởi con người) đôi khi là một nhiệm vụ khó khăn khi phải làm việc với một văn bản lớn, chứa nhiều thông tin (Trương Quốc Định & Nguyễn Quang Dũng, 2012).

Các giai đoạn trong tóm tắt văn bản: Theo quan điểm của các nhà nghiên cứu tóm tắt văn bản thì bản tóm tắt là một bản rút gọn của văn bản gốc thông qua việc lựa chọn và tổng quát hóa các khái niệm quan trọng (Hovy & Lin, 1999; Jones, 1999; Maybury & Mani, 2001). Hệ thống tóm tắt văn bản được chia làm ba giai đoạn chính:

(1) Phân tích (Analysis or Interpretation): Phân tích văn bản đầu vào để đưa ra những mô tả bao gồm các thông tin dùng để tìm kiếm, đánh giá các đơn vị ngữ liệu quan trọng cũng như các tham số đầu vào cho việc tóm tắt.

(2) Biến đổi (Transformation): Lựa chọn các thông tin trích chọn được biến đổi để giản lược và thống nhất, kết quả là các đơn vị ngữ liệu đã được tóm tắt.

(3) Tổng hợp (Sythesis or Realization): Từ các ngữ liệu đã tóm tắt, tạo văn bản mới chứa những điểm chính và quan trọng của văn bản gốc.



Hình 1. Các giai đoạn của hệ thống tóm tắt văn bản

Có hai phương pháp tóm tắt văn bản là: Tóm tắt trích chọn (Extractive), và tóm tắt trừu tượng (Abstractive).

(1) Phương pháp trích chọn: Tóm tắt trích chọn (Kiyoumarsi & Esfahani, 2011) được xây dựng bằng cách trích xuất các đơn vị văn bản quan trọng (câu hoặc đoạn văn) từ văn bản gốc, dựa trên phân tích từ/cụm từ, tần số, vị trí hoặc các từ gợi ý để xác định tầm quan trọng của các đơn vị, và từ đó trích xuất các đơn vị quan trọng nhất như là tóm tắt.

Để hiểu rõ hơn về cách thức hoạt động của các hệ thống tóm tắt loại trích chọn, nghiên cứu này mô tả ba nhiệm vụ khá độc lập mà tất cả các hệ thống tóm tắt trích chọn cần thực hiện: (1) Biến đổi văn bản, hay nói cách khác là dùng các thuật toán về thống kê, đồ thị hóa, máy học… để biểu diễn văn bản; (2) Tính trọng số về tính quan trọng của câu; và (3) Chọn một tập con trong văn bản gốc để trở thành văn bản tóm tắt. Trong nghiên cứu này, nhóm tác giả thực hiện tóm tắt văn bản theo phương pháp này.

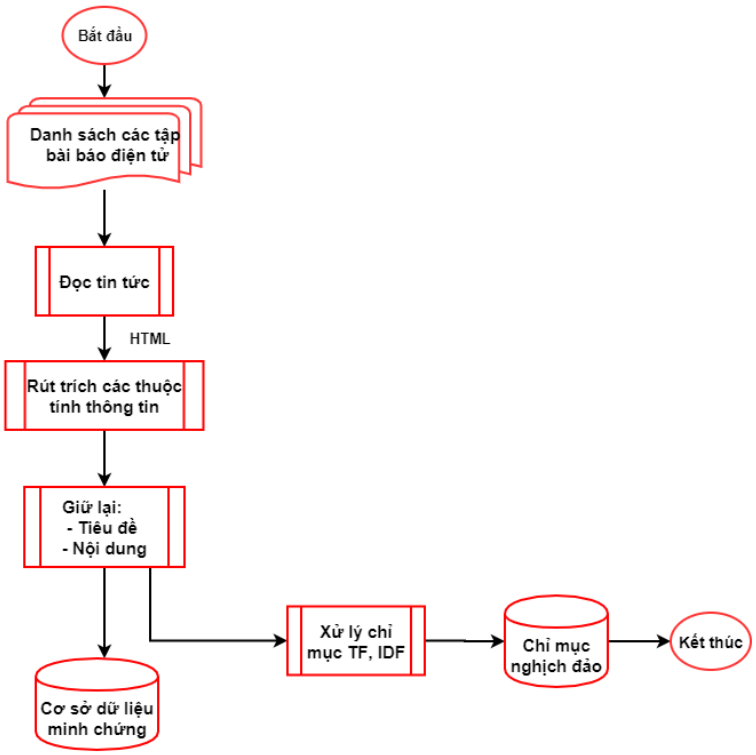
(2) Tóm tắt trừu tượng: Hướng tiếp cận tóm tắt trừu tượng (Erkan & Radev, 2004), có nghĩa là hệ thống cố gắng hiểu được ý chính của tài liệu rồi sau đó diễn giải chúng dưới dạng ngôn ngữ tự nhiên.

3. Phương pháp nghiên cứu

3.1. Mô hình tổng quan hệ thống

Hệ thống phát hiện tin giả trên mạng xã hội bằng ngôn ngữ tiếng Việt được thực hiện qua hai phân hệ: Phân hệ thu thập dữ liệu và phân hệ phát hiện tin giả.

Phân hệ thu thập dữ liệu chủ yếu thu thập các tin tức từ các trang web đáng tin cậy như: https://vnexpress.net/, https://vietnamnet.vn/, https://tuoitre.vn/... chỉ giữ lại tiêu đề và nội dung, bỏ qua các hình ảnh, video và thực hiện lập chỉ mục nghịch đảo.

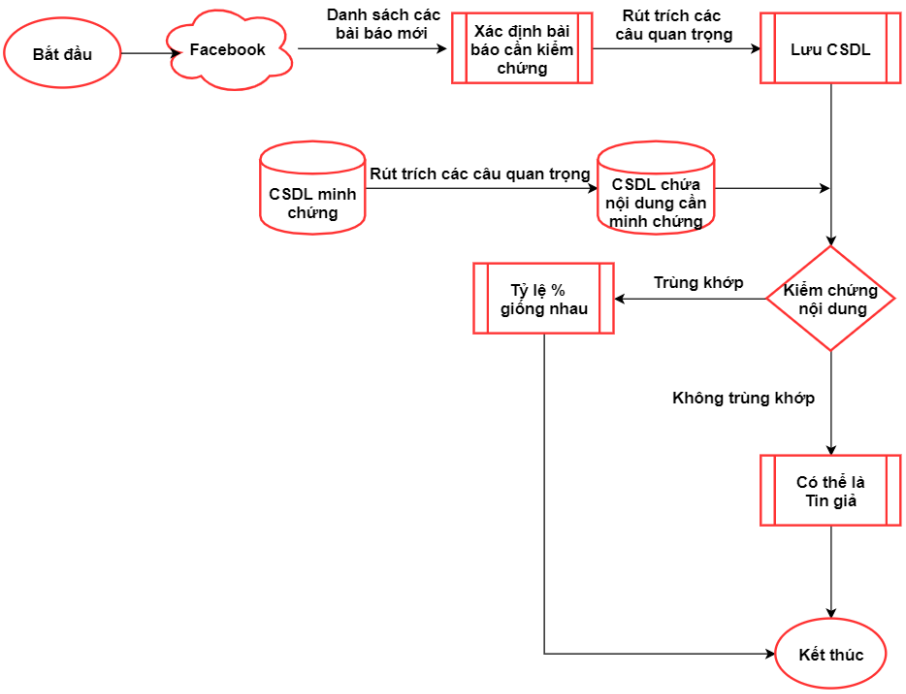


Hình 2. Phân hệ thu thập dữ liệu

Phân hệ thu thập dữ liệu được thực hiện qua các bước như sau:

- *Bước 1:* Thu thập các danh sách bài báo điện tử đáng tin cậy như: VnExpress, Vietnamnet, Thanh Niên, Tuổi Trẻ, Tin tức Online.
- *Bước 2:* Đọc tin tức.
- *Bước 3:* Dựa vào các mã HTML để rút trích các thuộc tính của tin tức.
- *Bước 4:* Dựa vào các mã HTML để giữ lại tiêu đề và nội dung của tin tức, bỏ qua hình ảnh và video.
- *Bước 5:* Lưu các trang tin tức lại để làm cơ sở dữ liệu minh chứng.
- *Bước 6:* Xử lý chỉ mục TF-IDF cho dữ liệu tin tức.
- *Bước 7:* Xây dựng chỉ mục nghịch đảo cho tin tức nhằm phục vụ cho việc tìm kiếm thông tin.

Phân hệ phát hiện tin giả thực hiện việc thu thập các tin từ mạng xã hội và kiểm tra xem tin đó có phải là tin giả hay không được thực hiện qua mô hình như Hình 3.



Hình 3. Phân hệ phát hiện tin giả

Các bước thực hiện ở phân hệ phát hiện tin giả như sau:

- *Bước 1:* Thu thập tin tức từ các trang Facebook như: Fanpage, group công khai.
- *Bước 2:* Xác định các bài báo cần kiểm chứng. Ở đây là các tin được nhiều người quan tâm có lượt xem, chia sẻ và bình luận cao.
- *Bước 3:* Rút trích các câu quan trọng từ các tin thu thập ở Bước 2.
- *Bước 4:* Từ cơ sở dữ liệu minh chứng đã lưu ở phần phân hệ thu thập dữ liệu, thực hiện rút trích các câu quan trọng.
- *Bước 5:* Thực hiện lưu cơ sở dữ liệu minh chứng.
- *Bước 6:* Thực hiện kiểm chứng nội dung giữa tin trên mạng xã hội và dữ liệu minh chứng nếu:

- Trùng khớp thì thể hiện tỷ lệ % giống nhau;

- Không trùng khớp thì thông báo có thể là tin giả.

3.2. Thu thập dữ liệu

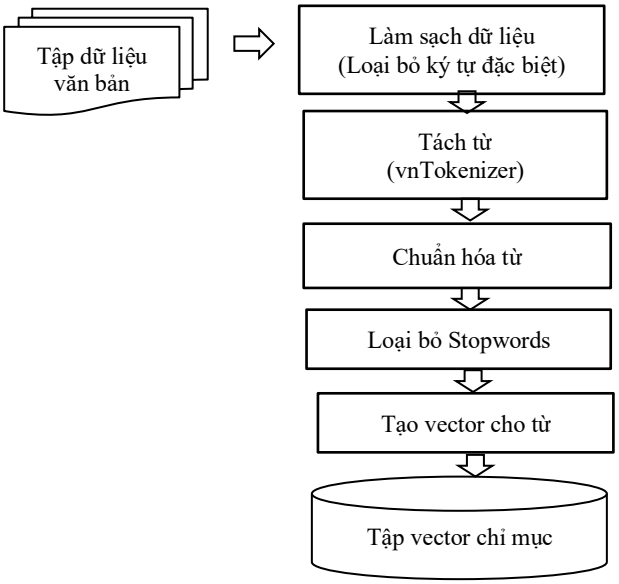
Để thực hiện phát hiện tin giả trên mạng xã hội, chúng ta có hai bộ dữ liệu để kiểm tra như sau:

- *Thứ nhất,* dữ liệu kiểm thử được thu thập từ nội dung của các bài viết từ các trang và group của các trang mạng xã hội như: Đại Kỷ Nguyên, SOHA, Đài Á Châu Tự Do, Người đưa tin, Nước Mỹ trở lại, RFI tiếng Việt, Chia sẻ cuộc sống… Đối với dữ liệu này, thu thập các tin có nhiều lượt xem, bình luận và chia sẻ.

- *Thứ hai,* bộ dữ liệu tin thật được thu thập từ các trang web đáng tin cậy như: https://vnexpress.net/, https://vietnamnet.vn/, https://thanhnien.vn/, https://tuoitre.vn/, https://www.24h.com.vn/, https://tintuconline.com.vn/. Việc lấy dữ liệu sử dụng RSS (Really Simple Syndication)[10] để gọi các trang web và phân tích các mã HTML, khối này chứa nội dung của tin như: Tiêu đề, nội dung tin và không chứa các thành phần khác như: Quảng cáo, menu, hình ảnh, video.

3.3. Tiền xử lý dữ liệu

Sau khi thu thập dữ liệu từ các trang web tin tức và các tin trên mạng xã hội theo cấu trúc nhất định, quá trình tiền xử lý dữ liệu sẽ được thực hiện. Đầu tiên chuyển dữ liệu về đúng dạng của nó và áp dụng các biện pháp tách từ để tách nội dung của văn bản thành các từ, cụm từ tương ứng, loại bỏ các ký tự dư thừa trong tiếng Việt, chỉ giữ lại các từ thật sự có ý nghĩa. Kết quả của giai đoạn tiền xử lý là các vector chỉ mục cho mỗi tài liệu văn bản. Các giai đoạn tiền xử lý được thực hiện qua Hình 4 như sau:



Hình 4. Sơ đồ tiền xử lý dữ liệu

3.4. Rút trích thông tin

Rút trích những thông tin quan trọng nhất từ một văn bản để tạo ra phiên bản ngắn gọn, súc tích nhưng vẫn chứa đựng đủ lượng thông tin cốt lõi của văn bản gốc với yêu cầu đảm bảo tính đúng đắn về ngữ pháp và chính tả.

Trong nghiên cứu này, nhóm tác giả áp dụng phương pháp rút trích văn bản tiếng Việt tự động được đề xuất bởi tác giả Trương Quốc Định và Nguyễn Quang Dũng (2012) để rút trích các câu quan trọng cho bộ dữ liệu kiểm thử thu thập từ mạng xã hội và bộ dữ liệu tin thật được thu thập từ các trang web đáng tin cậy.

[10] RSS: Tập tin có định dạng XML.

3.5. Xác định các đặc trưng

Để xác định các đặc trưng cho nội dung của bài viết, nghiên cứu sử dụng TF-IDF.

TF-IDF (Term Frequency - Inverse Document Frequency) là một phương thức thống kê được biết đến rộng rãi nhất để xác định độ quan trọng của một từ trong đoạn văn bản trong một tập nhiều đoạn văn bản khác nhau. Nó thường được sử dụng như một trọng số trong việc khai phá dữ liệu văn bản. TF-IDF chuyển đổi dạng biểu diễn văn bản thành dạng không gian vector (Vector Space Model – VSM), hoặc thành những vector thưa thớt (Nguyễn Việt Hùng, 2017).

TF (Term Frequency): Tần suất xuất hiện của một từ trong một đoạn văn bản được định nghĩa như sau:

$$TF(w)_D = 1 + log\frac{n_w(d)}{|d|} \quad (1)$$

IDF (Inverse Document Frequency): Tính toán độ quan trọng của một từ. Khi tính toán TF, mỗi từ đều quan trọng như nhau, nhưng có một số từ trong tiếng Anh như *"is", "of", "that"*... xuất hiện khá nhiều nhưng lại rất ít quan trọng. Vì vậy, chúng ta cần một phương thức bù trừ những từ xuất hiện nhiều lần và tăng độ quan trọng của những từ ít xuất hiện nhưng có ý nghĩa đặc biệt cho một số đoạn văn bản hơn bằng cách tính IDF:

$$TF(w)_D = 1 + log\frac{|D|}{|\{d:D|w \in d\}|} \quad (2)$$

TF-IDF từ w đối với tài liệu d và tổng số văn bản trong tập D được tính như sau:

$$TF - IDF(w)_{d,D} = TF(w)_d \times IDF(w)_D \quad (3)$$

3.6. Tính độ tương đồng

Nhằm xác định độ tương đồng giữa các tin trên mạng xã hội và tin thật được thu thập từ các trang tin tức, nghiên cứu sử dụng phương pháp tính độ tương đồng bằng cosine, các văn bản được biểu điễn theo mô hình không gian vector. Mỗi thành phần vector chỉ đến một từ tương ứng trong danh sách mục từ đã qua tiền xử lý ban đầu, các bước tiền xử lý gồm: Tách câu, tách từ, loại bỏ những từ không hợp lệ, tóm tắt văn bản, và biểu diễn câu trên không gian vector.

Không gian vector có kích thước bằng số mục từ trong danh sách mục từ chính. Mỗi phần tử là độ quan trọng của mục từ tương ứng trong câu. Độ quan trọng của từ j được tính bằng TF như sau:

$$w_{i,j} = \frac{tf_{i,j}}{\sqrt{\sum_j tf_{i,j}^2}} \quad (4)$$

Trong đó, $tf_{i,j}$: Tần số xuất hiện của mục từ i trong câu j.

Với không gian biểu diễn tài liệu được chọn là không gian vector và trọng số TF, độ đo tương đồng được chọn là cosine của góc giữa hai vector tương ứng của hai câu lần lượt là S_i và S_k. Vector biểu diễn hai câu lần lượt có dạng:

$S_i =< w_1^i, \ldots, w_t^i$ với w_t^i là trọng số của từ thứ t trong câu i;

$S_k =< w_1^k, \ldots, w_t^k$ với w_t^i là trọng số của từ thứ t trong câu k.

Độ tương tự giữa chúng được tính theo công thức:

$$Sim(S_i, S_k) = \frac{\sum_{j=1}^{t} w_j^i w_j^k}{\sqrt{\sum_{j=1}^{t}(w_j^i)^2 \times \sum_{j=1}^{t}(w_j^k)^2}} \quad (5)$$

4. Kết quả nghiên cứu

Trong nghiên cứu này, dữ liệu thực nghiệm gồm hai bộ dữ liệu là dữ liệu kiểm thử và dữ liệu tin thật. Dữ liệu kiểm thử được thu thập từ các trang mạng xã hội. Các tin trên Facebook được thu thập từ các trang và group công khai như: Đại Kỷ Nguyên, SOHA, Đài Á Châu Tự Do, Người đưa tin, Nước Mỹ trở lại, RFI tiếng Việt, Chia sẻ cuộc sống... với 100 tin thật và 200 tin giả. Dữ liệu tin thật thu thập từ các trang như: VnExpress, Vietnamnet, Thanh Niên, Tuổi Trẻ, 24h.com.vn, Tin tức Online với hơn 32.000 tin.

4.1. Xác định giá trị ngưỡng

Giá trị ngưỡng để kiểm tra tin giả hay tin thật được xác định dựa vào tính giá trị cosine của các tin có giá trị tương tự nhau được thu thập từ các trang gồm: VnExpress, Vietnamnet, Thanh Niên, Tuổi Trẻ, 24h.com.vn, Tin tức Online với 100 tin, ta có kết quả chạy thực nghiệm xác định giá trị ngưỡng giữa các tin như sau:

Bảng 1.

Kết quả chạy thực nghiệm từ tin thứ 1 đến tin thứ 10

DỮ LIỆU / KẾT QUẢ	1	2	3	4	5	6	7	8	9	10
Kết quả 1	0,54	0,37	0,27	0,19	0,45	0,20	0,39	0,19	0,53	0,30
Kết quả 2	0,61	0,30	0,24	0,11	0,35	0,20	0,23	0,19	0,52	0,25
Kết quả 3	0,48	0,51	0,10	0,27	0,53	0,12	0,74	0,19	0,53	0,81
Kết quả 4	0,20	0,37	0,17	0,10	0,44	0,25	0,62	0,28	0,54	0,68
Kết quả 5	0,28	0,38	0,13	0,22	0,30	0,17	0,23	0,17	0,47	0,63
Nhỏ nhất	0,20	0,30	0,10	0,10	0,30	0,12	0,23	0,17	0,47	0,25
Lớn nhất	0,61	0,51	0,27	0,27	0,53	0,25	0,74	0,28	0,54	0,81
Trung bình	0,42	0,39	0,18	0,18	0,41	0,19	0,44	0,20	0,52	0,53

Từ Bảng 1 cho thấy kết quả từ tin thứ 1 đến tin thứ 10 có giá trị cosine nhỏ nhất là 0,1; lớn nhất là 0,81; và giá trị trung bình là 0,35.

Bảng 2.

Kết quả chạy thực nghiệm từ tin thứ 11 đến tin thứ 20

DỮ LIỆU / KẾT QUẢ	11	12	13	14	15	16	17	18	19	20
Kết quả 1	0,44	0,6	0,44	0,78	0,18	0,41	0,26	0,24	0,45	0,15
Kết quả 2	0,28	0,05	0,42	0,72	0,23	0,21	0,25	0,18	0,18	0,16
Kết quả 3	0,53	0,64	0,46	0,77	0,11	0,54	0,34	0,24	0,18	0,31
Kết quả 4	0,27	0,25	0,22	0,71	0,16	0,11	0,30	0,12	0,63	0,18
Kết quả 5	0,32	0,67	1,00	0,65	0,07	0,65	0,10	0,34	0,48	0,10
Nhỏ nhất	0,27	0,05	0,22	0,65	0,07	0,11	0,10	0,12	0,18	0,10
Lớn nhất	0,53	0,67	1,00	0,78	0,23	0,65	0,34	0,34	0,63	0,31
Trung bình	0,37	0,44	0,51	0,73	0,15	0,38	0,25	0,22	0,38	0,18

Qua bảng kết quả chạy thực nghiệm từ tin thứ 11 đến tin thứ 20 cho thấy giá trị cosine nhỏ nhất là 0,05; cao nhất là 1; và giá trị trung bình là 0,36.

Bảng 3.

Kết quả chạy thực nghiệm từ tin thứ 21 đến tin thứ 30

DỮ LIỆU / KẾT QUẢ	21	22	23	24	25	26	27	28	29	30
Kết quả 1	0,43	0,18	0,47	0,33	0,21	0,18	0,33	0,32	0,50	0,37
Kết quả 2	0,29	0,27	0,52	0,38	0,28	0,37	0,13	0,15	0,57	0,24
Kết quả 3	0,34	0,28	0,34	0,35	0,19	0,18	0,23	0,12	0,33	0,37
Kết quả 4	0,22	0,26	0,27	0,18	0,18	0,25	0,14	0,06	0,42	0,32
Kết quả 5	0,18	0,31	0,23	0,59	0,21	0,16	0,28	0,08	0,45	0,99
Nhỏ nhất	0,18	0,18	0,23	0,18	0,18	0,16	0,13	0,06	0,33	0,24
Lớn nhất	0,43	0,31	0,52	0,59	0,28	0,37	0,33	0,32	0,57	0,99
Trung bình	0,29	0,26	0,37	0,37	0,21	0,23	0,22	0,15	0,45	0,46

Bảng 3 từ tin thứ 21 đến tin thứ 30 cho thấy giá trị cosine nhỏ nhất là 0,06; cao nhất là 0,99; và giá trị trung bình là 0,30.

Bảng 4.

Kết quả chạy thực nghiệm từ tin thứ 31 đến tin thứ 40

DỮ LIỆU / KẾT QUẢ	31	32	33	34	35	36	37	38	39	40
Kết quả 1	0,30	0,54	0,48	0,21	0,40	0,11	0,18	0,21	0,21	0,27
Kết quả 2	0,20	0,42	0,39	0,12	0,32	0,10	0,34	0,18	0,45	0,18
Kết quả 3	0,21	0,31	0,17	0,10	0,29	0,15	0,31	0,35	0,38	0,17
Kết quả 4	0,22	0,30	0,46	0,18	0,24	0,11	0,24	0,36	0,42	0,17
Kết quả 5	0,12	0,47	0,20	0,14	0,18	0,20	0,36	0,33	0,19	0,21
Nhỏ nhất	0,12	0,30	0,17	0,10	0,18	0,10	0,18	0,18	0,19	0,17
Lớn nhất	0,30	0,54	0,48	0,21	0,40	0,20	0,36	0,36	0,45	0,27
Trung bình	0,21	0,41	0,34	0,15	0,29	0,13	0,29	0,29	0,33	0,20

Bảng 4 kết quả chạy thực nghiệm từ tin thứ 21 đến tin thứ 30 hiển thị giá trị cosine nhỏ nhất là 0,1; cao nhất là 0,48; và giá trị trung bình là 0,26.

Bảng 5.

Kết quả chạy thực nghiệm từ tin thứ 41 đến tin thứ 50

DỮ LIỆU / KẾT QUẢ	41	42	43	44	45	46	47	48	49	50
Kết quả 1	0,38	0,40	0,17	0,05	0,44	0,64	0,42	0,63	0,43	0,07
Kết quả 2	0,22	0,37	0,23	0,20	0,60	0,48	0,25	0,60	0,71	0,14
Kết quả 3	0,16	0,26	0,19	0,35	0,38	0,69	0,32	0,07	0,65	0,26
Kết quả 4	0,11	0,49	0,11	0,29	0,50	0,59	0,26	0,67	0,53	0,25
Kết quả 5	0,26	0,57	0,46	0,10	0,26	0,50	0,52	0,44	0,20	0,19
Nhỏ nhất	0,11	0,26	0,11	0,05	0,26	0,48	0,25	0,07	0,20	0,07
Lớn nhất	0,38	0,57	0,46	0,35	0,60	0,69	0,52	0,67	0,71	0,26
Trung bình	0,23	0,42	0,23	0,20	0,44	0,58	0,35	0,48	0,50	0,18

Kết quả chạy thực nghiệm từ tin thứ 41 đến tin thứ 50 cho thấy giá trị cosine nhỏ nhất trong bảng là 0,07; lớn nhất là 0,71; và giá trị trung bình là 0,36.

Bảng 6.

Kết quả chạy thực nghiệm từ tin thứ 51 đến tin thứ 60

DỮ LIỆU / KẾT QUẢ	51	52	53	54	55	56	57	58	59	60
Kết quả 1	0,38	0,44	0,21	0,15	0,38	0,52	0,26	0,34	0,26	0,15
Kết quả 2	0,34	0,15	0,25	0,18	0,36	0,49	0,23	0,31	0,27	0,42
Kết quả 3	0,32	0,34	0,25	0,16	0,25	0,47	0,30	0,49	0,41	0,22
Kết quả 4	0,25	0,56	0,09	0,12	0,33	0,37	0,22	0,45	0,10	0,22
Kết quả 5	0,28	0,12	0,19	0,06	0,16	0,34	0,11	0,30	0,13	0,18
Nhỏ nhất	0,25	0,12	0,09	0,06	0,16	0,34	0,11	0,30	0,10	0,15
Lớn nhất	0,38	0,56	0,25	0,18	0,38	0,52	0,30	0,49	0,41	0,42
Trung bình	0,31	0,32	0,20	0,13	0,30	0,44	0,22	0,38	0,23	0,24

Kết quả Bảng 6 hiển thị các giá trị cosine từ tin thứ 51 đến tin thứ 60, trong đó, giá trị nhỏ nhất là 0,09; lớn nhất là 0,56; và giá trị trung bình là 0,28.

Bảng 7.

Kết quả chạy thực nghiệm từ tin thứ 61 đến tin thứ 70

DỮ LIỆU / KẾT QUẢ	61	62	63	64	65	66	67	68	69	70
Kết quả 1	0,14	0,30	0,22	0,30	0,23	0,22	0,22	0,59	0,09	0,43
Kết quả 2	0,12	0,22	0,33	0,53	0,06	0,18	0,18	0,27	0,09	0,30
Kết quả 3	0,09	0,27	0,42	0,34	0,09	0,16	0,16	0,55	0,13	0,37
Kết quả 4	0,05	0,16	0,40	0,41	0,11	0,47	0,47	0,32	0,19	0,30
Kết quả 5	0,11	0,33	0,38	0,49	0,13	0,19	0,19	0,18	0,03	0,26
Nhỏ nhất	0,05	0,16	0,22	0,30	0,06	0,16	0,16	0,18	0,03	0,26
Lớn nhất	0,14	0,33	0,42	0,53	0,23	0,47	0,47	0,59	0,19	0,43
Trung bình	0,10	0,26	0,35	0,41	0,12	0,24	0,24	0,38	0,11	0,33

Bảng 7 hiển thị các giá trị cosine từ tin thứ 61 đến tin thứ 70, trong đó, giá trị nhỏ nhất là 0,05; lớn nhất là 0,59; và giá trị trung bình là 0,26.

Bảng 8.

Kết quả chạy thực nghiệm từ tin thứ 71 đến tin thứ 80

DỮ LIỆU / KẾT QUẢ	71	72	73	74	75	76	77	78	79	80
Kết quả 1	0,14	0,78	0,80	0,18	0,24	0,17	0,22	0,20	0,11	0,29
Kết quả 2	0,22	0,61	0,74	0,13	0,32	0,14	0,17	0,08	0,13	0,32
Kết quả 3	0,52	0,75	0,73	0,04	0,19	0,16	0,22	0,08	0,17	0,42
Kết quả 4	0,38	0,59	0,60	0,60	0,16	0,15	0,20	0,25	0,12	0,06
Kết quả 5	0,21	0,35	0,29	0,29	0,20	0,16	0,26	0,14	0,16	0,26
Nhỏ nhất	0,14	0,35	0,29	0,04	0,16	0,14	0,17	0,08	0,11	0,06
Lớn nhất	0,52	0,78	0,80	0,60	0,32	0,17	0,26	0,25	0,17	0,42
Trung bình	0,29	0,62	0,63	0,25	0,22	0,16	0,21	0,15	0,14	0,27

Kết quả Bảng 8 hiển thị giá trị cosine nhỏ nhất là 0,04; lớn nhất là 0,8; và giá trị trung bình từ tin thứ 71 đến tin thứ 80 là 0,29.

Bảng 9.

Kết quả chạy thực nghiệm từ tin thứ 81 đến tin thứ 90

DỮ LIỆU / KẾT QUẢ	81	82	83	84	85	86	87	88	89	90
Kết quả 1	0,32	0,22	0,27	0,15	0,35	0,24	0,26	0,27	0,23	0,40
Kết quả 2	0,15	0,41	0,24	0,35	0,18	0,13	0,14	0,38	0,14	0,26
Kết quả 3	0,32	0,55	0,23	0,14	0,34	0,21	0,08	0,23	0,12	0,30
Kết quả 4	0,30	0,31	0,41	0,13	0,20	0,18	0,12	0,35	0,08	0,25
Kết quả 5	0,55	0,25	0,35	0,15	0,20	0,10	0,12	0,31	0,07	0,25
Nhỏ nhất	0,15	0,22	0,23	0,13	0,18	0,10	0,08	0,23	0,07	0,25
Lớn nhất	0,55	0,55	0,41	0,35	0,35	0,24	0,26	0,38	0,23	0,4
Trung bình	0,33	0,35	0,30	0,18	0,25	0,17	0,14	0,31	0,13	0,29

Kết quả các giá trị trong Bảng 9 có giá trị cosine nhỏ nhất là 0,07; cao nhất là 0,55; và giá trị trung bình từ tin thứ 81 đến tin thứ 90 là 0,25.

Bảng 10.

Kết quả chạy thực nghiệm từ tin thứ 91 đến tin thứ 100

DỮ LIỆU / KẾT QUẢ	91	92	93	94	95	96	97	98	99	100
Kết quả 1	0,54	0,20	0,29	0,38	0,41	0,57	0,51	0,18	0,26	0,33
Kết quả 2	0,44	0,13	0,14	0,22	0,30	0,10	0,38	0,19	0,23	0,26
Kết quả 3	0,26	0,09	0,11	0,40	0,46	0,36	0,36	0,12	0,19	0,27
Kết quả 4	0,45	0,12	0,10	0,74	0,29	0,82	0,30	0,17	0,17	0,32
Kết quả 5	0,45	0,10	0,05	0,49	0,15	0,65	0,17	0,20	0,07	0,18
Nhỏ nhất	0,26	0,09	0,05	0,22	0,15	0,10	0,17	0,12	0,07	0,18
Lớn nhất	0,54	0,20	0,29	0,74	0,46	0,82	0,51	0,19	0,26	0,33
Trung bình	0,43	0,13	0,14	0,45	0,32	0,50	0,34	0,16	0,18	0,27

Các kết quả trong Bảng 10 có giá trị cosine nhỏ nhất là 0,09; lớn nhất là 0,82; và giá trị trung bình từ tin thứ 91 đến tin thứ 100 là 0,25.

Từ các kết quả ở Bảng 10 để xác định giá trị ngưỡng kiểm tra tin giả, tin thật và có thể tin giả. Nhóm nghiên cứu thực hiện phép tính trung bình các giá trị từ Bảng 1 đến Bảng 10, ta được ngưỡng để kiểm tra tin trong bài báo này là 0,3.

4.2. Kiểm tra kết quả

Với tập dữ liệu kiểm thử được thu thập 100 tin thật và 200 tin giả được gán nhãn bằng tay và thực hiện kiểm chứng bằng phần mềm, đánh giá mô hình tin giả và tin thật, ta có các kết quả như sau:

Bảng 11.

Kết quả kiểm tra tin giả, tin thật bằng phần mềm

STT	Nhãn	Số mẫu	Kết quả kiểm tra bằng phần mềm	
			Đúng	Sai
1	Tin giả	200	176	24
2	Tin thật	100	34	66

Bảng 11 trình bày kết quả kiểm tra tin giả và tin thật khi chạy bằng phần mềm đối với 200 tin giả. Khi sử dụng phần mềm kiểm tra cho kết quả đúng với tin giả là 176 tin và 24 tin cho kết quả sai. Các tin sai này chủ yếu thuộc các chủ đề: Thế giới, sức khỏe.

Sau khi kiểm tra kết quả tin giả và tin thật bằng phần mềm, nhóm tác giả thực hiện đánh giá mô hình qua các độ đo Precision[11], Recall[12], F1-Score[13], Accuracy[14].

Bảng 12.

Kết quả đánh giá mô hình tin giả, tin thật

STT	Nhãn	Precison	Recall	F1-Score	Accuracy
1	Tin giả	50%	44%	47%	88%
2	Tin thật	50%	17%	25%	34%

Ở Bảng 12 cho thấy giá trị Precison của tin giả và tin thật như nhau là 50%, giá trị Recall, F1-Score, Accuracy của tin giả cao hơn tin thật từ 22% đến 54%. Từ đó, chúng ta thấy rằng các tin giả có độ chính xác cao hơn tin thật.

Bảng 13.

Kết quả đánh giá toàn hệ thống

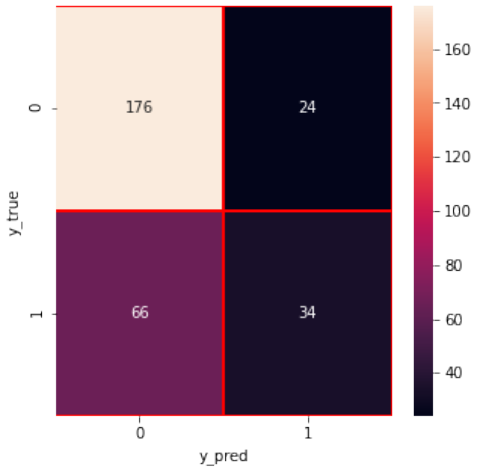
STT	Mô hình đánh giá	Tỷ lệ %
1	Precision	66
2	Recall	61
3	F1-Score	61
4	Accuracy	70

Kết quả trong Bảng 13 cho thấy giá các giá trị Precison, Recall, F1-Score đều trên 60%, giá trị Accuracy tương đối khá cao khoảng 70%. Do đó, ta thấy rằng hệ thống có thể kiểm tra tương đối chính xác tin trên mạng xã hội có phải là tin giả hay không.

Ngoài các độ đo trên, nhóm tác giả sử dụng Confustion Matrix[15] để đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp với các giá trị như sau:

- True Positive (TP): Dự đoán các mẩu tin tức giả được chú thích là tin tức giả: 176;
- True Negative (TN): Dự đoán các mẩu tin tức thật được chú thích là tin tức thật: 34;
- False Negative (FN): Dự đoán các mẩu tin tức thật được chú thích là tin giả: 66;
- False Positive (FP): Dự đoán các mẩu tin tức giả mà được chú thích là tin thật: 24.

[11] Precision: Độ chính xác.
[12] Recall: Tỷ lệ trọng số điểm mô hình dự đoán đúng trên tổng số điểm dự đoán đúng ban đầu.
[13] F1-Score: Trung bình của Precision và Recall.
[14] Accuracy: Độ chuẩn xác.
[15] Confustion Matrix : Ma trận nhầm lẫn hay ma trận lỗi



Hình 5. Confustion Matrix

5. Kết luận

Nghiên cứu xây dựng thành công giải pháp phát hiện tin giả trên mạng xã hội bằng ngôn ngữ tiếng Việt. Với tập dữ liệu tin thật được thu thập từ các trang: VnExpress, Vietnamnet, Tuổi Trẻ, Thanh Niên, Tin tức Online, 24h.com.vn với hơn 32.000 tin và tập dữ liệu tin kiểm chứng thu thập từ các trang và group trên mạng xã hội. Trong nghiên cứu này, thu thập từ Facebook với 100 tin thật và 200 tin giả bằng ngôn ngữ tiếng Việt. Nghiên cứu thực hiện kiểm chứng tin giả thành công bằng phương pháp cosine với độ chính xác là 70%. Nghiên cứu có giá trị tham khảo cho các ứng dụng phát hiện tin giả trên mạng xã hội bằng ngôn ngữ tiếng Việt. Tuy nhiên, nghiên cứu này vẫn còn một số hạn chế, có thể tiếp tục tiếp tục thực hiện trong thời gian tới:

- *Thứ nhất,* về thu thập dữ liệu, nghiên cứu này chỉ thu thập các tin trên mạng xã hội Facebook có thể mở rộng trên các trang Twitter và Instagram.

- *Thứ hai,* về phương pháp nghiên cứu, nghiên cứu này chỉ sử dụng phương pháp tính độ tương đồng cosine, nếu kết hợp phương pháp tính độ tương đồng theo đồ thị so sánh các đỉnh (Graph Vertices Comparison) (Truong và cộng sự, 2008) sẽ cho kết quả tốt hơn.

- *Thứ ba,* do tính phức tạp của tin giả nên giải pháp chỉ đề xuất một tin trên mạng xã hội là tin giả chứ chưa khẳng định là tin giả. Những định hướng nghiên cứu tiếp theo sẽ tập trung vào việc phân tích ngữ nghĩa của bài viết và phân loại theo chủ đề như: Thời sự, thế giới, giải trí, sức khỏe, thể thao, pháp luật.

Tài liệu tham khảo

Ahmed, H. (2017). *Detecting opinion spam and fake news using N-gram analysis and semantic similarity.* A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Applied Science. University of Victoria.

Aldwairi, M., & Alwahedi, A. (2018). Detecting fake news in social media networks. *Procedia Computer Science*, *141*, 215–222.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.

Bông Mai. (2017). *Tin tức giả, hệ quả thật*. Truy cập ngày 30/12/2017, từ https://nhandan.vn/megastory/2017/12/29/

Burfoot, C., & Baldwin, T. (2009). *Automatic satire detection: Are you having a laugh?*. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 161–164). Suntec, Singapore. Association for Computational Linguistics.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Horne, B. D., & Adali, S. (2017). *This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news*. In Proceedings of the International AAAI Conference on Web and Social Media, *11*(1), 759–766.

Hovy, E., & Lin, C. Y. (1999). Automated text summarization in SUMMARIST. In *Advances in Automatic Text Summarization* (pp. 81–94). Cambridge, MA: The MIT Press.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). *News verification by exploiting conflicting social viewpoints in microblogs.* In Proceedings of the AAAI Conference on Artificial Intelligence, *30*(1). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/10382

Jones, K. S. (1999). Automatic summarizing: Factors and directions. In *Advances in Automatic Text Summarization* (pp. 1–12). Cambridge, MA: The MIT Press.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). *Prominent features of rumor propagation in online social media.* In 2013 IEEE 13th International Conference on Data Mining (pp. 1103–1108), 7–10 December, 2013, Dallas, TX, USA.

Kiyoumarsi, F., & Esfahani, F. R. (2011). *Optimizing Persian text summarization based on fuzzy logic.* In 2011 International Conference on Intelligent Building and Management (pp. 264–269). Singapore: IACSIT Press.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B.,..., & Han, J. (2016). A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, *17*(2), 1–16.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F., & Cha, M. (2016). *Detecting rumors from microblogs with recurrent neural networks.* In 25th International Joint Conference on Artificial Intelligence, IJCAI 2016, 9–15 July, 2016, New York, United States.

Maybury, M. T., & Mani, I. (2001). *Automatic summarization.* In American/European Conference on Computational Linguistics, 8 July, 2001, Toulouse, France.

Mukherjee, S., & Weikum, G. (2015). *Leveraging joint interactions for credibility analysis in news communities.* In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 353–362), 19–23 October, 2015, Melbourne, VIC, Australia.

Nguyễn Việt Hùng. (2017). *Trích chọn thuộc tính trong đoạn văn bản với TF-IDF.* Truy cập ngày 20/10/2017, từ https://viblo.asia/p/trich-chon-thuoc-tinh-trong-doan-van-ban-voi-tf-idf-Az45bAOqlxY

Ruchansky, N., Seo, S., & Liu, Y. (2017). *CSI: A hybrid deep model for fake news detection*. In Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM) 2017 (pp. 797–806).

Sampson, J., Morstatter, F., Wu, L., & Liu, H. (2016*). Leveraging the implicit structure within social media for emergent rumor detection.* In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 2377–2382), 24–28 October, 2016, Indianapolis, IN, USA.

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.

Thanh Hà. (2018). *Tin giả hoành hành, Luật Việt Nam xử phạt như thế nào?*. Truy cập ngày 06/05/2018, từ htttps://tuoitre.vn/tin-gia-hoanh-hanh-luat-viet-nam-xu-phat-the-nao 20180506080658065.htm

Thủ tướng Chính phủ Việt Nam. (2020). *Nghị định số 15/2020/NĐ-CP của Chính phủ: Quy định xử phạt vi phạm hành chính trong lĩnh vực bưu chính, viễn thông, tần số vô tuyến điện, công nghệ thông tin và giao dịch điện tử*, ban hành ngày 03/02/2020. Truy cập từ https://vanban.chinhphu.vn/default.aspx?pageid=27160&docid=199053

Trương Quốc Định, & Nguyễn Quang Dũng. (2012). *Một giải pháp tóm tắt văn bản tiếng Việt tự động*. Hội thảo quốc gia lần thứ XV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, tổ chức ngày 12/2012, tại Hà Nội.

Truong, Q. D., Dkaki, T., Mothe, J., & Charrel, P. J. (2008). *GVC: A graph-based information retrieval mode.* In Conférence en Recherche d'Information et Applications *CORIA* (pp. 337–351).

Vương Thị Hồng, Nguyễn Minh Đức, Nguyễn Văn Quang, Trần Văn Hiến, Nguyễn Thị Cẩm Vân, & Trần Mai Vũ. (2016). *Phát hiện tài khoản spam trên mạng xã hội dựa trên phương pháp lai (Một thực nghiệm)*. Hội thảo lần thứ I: Một số vấn đề chọn lọc về an toàn an ninh thông tin, tổ chức ngày 28/11/2016, tại Hà Nội.

Weikum, G. (2017). *What computers should know, shouldn't know, and shouldn't believe*. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 1559–1560). doi: 10.1145/3041021.3051120

Wu, L., Li, J., Hu, X., & Liu, H. (2017). *Gleaning wisdom from the past: Early detection of emerging rumors in social media.* In Proceedings of the 17th SIAM International Conference on Data Mining (pp. 99–107). Houston, United States: Society for Industrial and Applied Mathematics.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, *51*(2), 1–36.