MỘT MÔ HÌNH ĐA PHƯƠNG PHÁP CHO PHÁT HIỆN TIN BÀI PHẢN ĐỘNG TIẾNG VIỆT

Lê ngọc An[*], Hoàng Xuân Dậu[#], Dương Trần Đức[+], Đinh Tuấn Long[*]
[*]Khoa Công nghệ thông tin, Trường Đại học Mở Hà Nội
[#]Khoa An toàn thông tin, Học viện Công nghệ Bưu chính Viễn thông
[+]Khoa Công nghệ Thông tin 1, Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Các dạng tin giả, tin bài có nội dung độc hại, phản động hiện nay được đăng tải và lan truyền rất mạnh do sự phổ biến của mạng Internet và đặc biệt là sự bùng nổ của các mạng xã hội, các dịch vụ trực tuyến trên không gian mạng. Các tin bài có nội dung độc hại, và đặc biệt là các tin bài phản động nhắm đến nước ta, như các tin bài tung tin thất thiệt, nói xấu lãnh tụ, kích động phá hoại khối đại đoàn kết toàn dân tộc có ảnh rất lớn đến đời sống xã hội do chúng khả năng lan truyền nhanh và có nhiều hình thức thể hiện, như tin bài dưới dạng văn bản, ảnh, hoặc kết hợp. Do sự nghiêm trọng của các bài viết đăng tin giả, hoặc có nội dung độc hại, phản động trên không gian mạng, đã có một số nghiên cứu ở trong và ngoài nước cho phát hiện và phòng chống. Tuy vậy, đa số các đề xuất tập trung xử lý tin bài có nội dung giả mạo, độc hại được đăng tải sử dụng ngôn ngữ tiếng Anh. Hơn nữa, do một số lượng lớn tin bài được đăng tải dưới dạng hình ảnh, hoặc văn bản nhúng trong ảnh, video, nên việc xử lý gặp nhiều khó khăn, dẫn đến tỷ lệ phát hiện đúng còn tương đối thấp. Bài báo này đề xuất một mô hình đa phương pháp dựa trên sự kết hợp của các mô hình PhoBERT và Swin Transformer V2 cho phát hiện tin bài phản động dưới dạng văn bản và hình ảnh. Kết quả thử nghiệm cho thấy mô hình kết hợp đề xuất sử dụng đặc trưng văn bản và ảnh cho các hiệu suất phát hiện vượt trội so với các mô hình riêng lẻ và các mô hình đã có, với các độ đo Accuracy đạt 97%, Precision đạt 97%, Recall đạt 97.5% và F1-score đạt 97%.

Từ khóa: Phát hiện tin bài phản động, mô hình đa phương pháp, PhoBERT, Swin Transformer V2, Swin Transformer v2 + PhoBERT.

I. GIỚI THIỆU

Sự phổ biến của mạng Internet ngày càng giúp cho việc giao tiếp, trao đổi thông tin trở nên thuận tiện và dễ dàng hơn bao giờ hết. Bên cạnh đó, việc các tin tức lan truyền quá nhanh mà không được kiểm soát đầy đủ cũng đem đến những thách thức lớn, đặc biệt là vấn đề về tin tức có nội dung sai sự thật, tin giả, độc hại, hoặc phản động. Trong hoàn cảnh hiện nay, các thế lực thù địch trong và ngoài nước liên kết với nhau đẩy mạnh các hoạt động chống phá cách mạng nước ta dưới nhiều hình thức, cả về chính trị, kinh tế và văn hóa. Chẳng hạn, chúng tung tin thất thiệt, nói xấu lãnh tụ; kích động phá hoại khối đại đoàn kết toàn dân tộc; chia rẽ giữa Đảng với nhân dân,... Các hoạt động chống phá bằng cách lan truyền tin bài có nội dung độc hại, phản động đã làm cho một bộ phận nhân dân, nhất là lớp trẻ, đồng bào dân tộc thiểu số hoang mang, dao động, từ đó tạo sự hoài nghi, bất mãn với chế độ. Theo Báo điện tử của Đảng Cộng sản Việt nam, trung bình một tháng, các thế lực thù địch phát tán hơn 130.000 bài viết, video xuyên tạc lên các mạng xã hội và các nền tảng khác trên Internet, trong đó số lượng tin giả, xấu độc chiếm trên 50% [1]. Theo đó, có hơn 80.000 bài viết được phát tán trên mạng xã hội Facebook, chiếm 67% và khoảng 40.000 bài viết, video xuyên tạc được đăng tải trên các kênh của mạng xã hội Youtube, các blog cá nhân hoặc các trang tin tức phản động khác. Lợi dụng Internet và đặc biệt là các mạng xã hội, các thế lực thù địch đã lập hàng nghìn trang tin, blog, hàng trăm tờ báo, nhà xuất bản và các đài phát thanh truyền hình có chương trình tiếng Việt để xuyên tạc, nói xấu Đảng Cộng sản, chế độ xã hội chủ nghĩa ở Việt Nam.

Nghiên cứu của Cao và cộng sự [2] cho thấy các bài có video, hoặc hình ảnh nhận được nhiều hơn 18% lượt nhấp chuột, 89% lượt thích, và 150% lượt chia sẻ lại so với các bài đăng không có video. Để thu hút người xem, các tin, bài phản động được đăng tải thường có nội dung kích động, giật gân, kích thích tính tò mò của người xem. Theo trang tin VTV.vn, ngày 11/6/2023 đã xảy ra một vụ khủng bố nghiêm trọng tại tỉnh Đắk Lắk làm 9 người chết, 2 người bị thương, trụ sở cùng nhiều trang thiết bị của chính quyền 2 xã bị đốt phá mà nguyên nhân sâu xa là do người dân tin vào những tin, bài phản động, chống chính quyền của thế lực thù địch ở nước ngoài . Đây là một minh chứng cho những ảnh hưởng tiêu cực của các tin, bài có nội dung độc hại, phản động. Do vậy, việc nghiên cứu phát hiện tin, bài có nội dung độc hại, phản động trên không gian mạng, đặc biệt là trên các mạng xã hội là việc làm cấp thiết, có tính thực tiễn hiện nay.

Do tính chất nghiêm trọng của các bài viết đăng tin giả, hoặc có nội dung độc hại, phản động trên không gian mạng, đã có một số nghiên cứu ở trong và ngoài nước cho phát hiện và phòng chống, như [3], [4], [5], [6], [7]. Mặc dù vậy, đa số các nghiên cứu tập trung xử lý tin, bài có nội dung giả mạo, hoặc độc hại được đăng tải sử dụng ngôn ngữ tiếng Anh. Hơn nữa, do một số lượng lớn tin bài được đăng tải dưới dạng hình ảnh, hoặc văn bản nhúng trong ảnh, video, hoặc kết hợp giữa nội dung văn bản và ảnh, nên việc xử lý gặp nhiều khó khăn, dẫn đến tỷ lệ phát hiện đúng còn tương

Tác giả liên hệ: Hoàng Xuân Dậu,
Email: dauhx@ptit.edu.vn
Đến tòa soạn: 10/2024, chỉnh sửa: 11/2024, chấp nhận đăng: 12/2024.

đối thấp và tỷ lệ cảnh báo sai còn cao. Bài báo này đề xuất mô hình phát hiện tin bài phản động tiếng Việt dựa trên kết hợp mô hình PhoBERT [8] và mô hình Swin Transformmer V2 [9] nhằm xử lý hiệu quả hai dạng tin bài phản động phổ biến, bao gồm tin bài dưới dạng ảnh và văn bản. Mô hình đề xuất có khả năng phân biệt tin bài phản động với tin bài bình thường tốt hơn nhờ sử dụng mô hình PhoBERT cho nhận diện các đặc trưng của văn bản và mô hình Swin Transformmer V2 cho nhận diện các đặc trưng của hình ảnh trong tin bài. Các đóng góp chính của bài báo gồm:

- Đề xuất mô hình phát hiện tin bài phản động tiếng Việt dựa trên sự kết hợp của mô hình PhoBERT và mô hình Swin Transformmer V2;

- Thu thập tập dữ liệu tin bài phản động tiếng Việt, thử nghiệm và đánh giá mô hình phát hiện tin bài phản động tiếng Việt đề xuất.

Phần còn lại của bài báo được cấu trúc như sau: Phần II trình bày một số nghiên cứu liên quan trong lĩnh vực phân loại văn bản và bài báo. Phần III mô tả phương pháp. Phần IV trình bày về các kết quả và thảo luận. Cuối cùng, các kết luận sẽ được trình bày trong phần V của bài báo.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Mục này giới thiệu một số nghiên cứu có liên quan gần đến mô hình phát hiện tin bài phản động đề xuất trong bài báo, bao gồm Nguyen và Gokhale [3], Uppada và cộng sự [4], Armin và các cộng sự [5], Wu và cộng sự [6] và Kiela và cộng sự [13]. Trong đó, Nguyen và Gokhale [3] đề xuất mô hình nhận diện ý kiến chống chính phủ trên Twitter sử dụng ngôn ngữ tiếng Anh trong các cuộc biểu tình chống phong tỏa có động cơ chính trị tại thủ đô của bang Michigan, Mỹ. Mô hình sử dụng các kỹ thuật n-grams và TF-IDF để tách và tính giá trị cho các đặc trưng từ các bài viết. Các tác giả sử dụng các thuật toán, như Random Forests, SVM, Logistic Regression, DistilBERT, MLP để xây dựng các bộ phân loại. Các bộ phân loại có thể phát hiện hiệu quả các bài viết có quan điểm chống chính phủ với độ chính xác khoảng 85% và độ đo F1 khoảng 82%. Nhược điểm của mô hình này là chỉ xử lý được các bài viết dạng văn bản mà chưa xử lý được những bài viết dưới dạng ảnh. Ngoài ra, mô hình được phát triển cho xử lý các bài viết tiếng Anh và được giới hạn trong bối cảnh cụ thể của các cuộc biểu tình chống phong tỏa tại bang Michigan. Điều này có nghĩa là hiệu suất phân loại của mô hình có thể bị giảm khi áp dụng cho các bài viết sử dụng ngôn ngữ khác hoặc trong các bối cảnh khác.

Theo một hướng tiếp cận khác, Uppada và cộng sự [4] đề xuất một mô hình cho phép nhận diện các bài viết có nội dung giả mạo. Mô hình này sử dụng dữ liệu các bài viết trích từ tập dữ liệu Fakeddit có ngôn ngữ là tiếng Anh, với hơn 1 triệu mẫu chứa dữ liệu văn bản, hình ảnh, siêu dữ liệu và tiêu đề được thu thập từ nhiều nguồn khác nhau. Nhóm tác giả đã sử dụng BERT+Dense và RoBERTa+Dense cho xử lý ngôn ngữ tự nhiên từ các bài viết, sau đó thực hiện so sánh các mô hình xử lý, bao gồm Xception, Inception-ResNet-V2, ResNet50, VGG19. Kết quả thử nghiệm cho thấy, mô hình cho hiệu suất phát hiện tốt nhất là mô hình kết hợp Xception + (BERT + Dense), với Accuracy đạt 91,94%, Precision đạt 93,43%, Recall đạt 93,07% và F1-score đạt 93%. Hạn chế của mô hình là chỉ được huấn luyện trên dữ liệu tiếng Anh, nên hiệu suất phát hiện có thể giảm khi áp dụng cho xử lý các bài viết trên các ngôn ngữ khác, đặc biệt là tiếng Việt.

Armin và cộng sự [5] đã đưa ra một phương pháp phát hiện thông tin sai lệch trên mạng xã hội bằng cách sử dụng nhiều loại dữ liệu khác nhau có ngôn ngữ là tiếng Anh trên văn bản, hình ảnh, bình luận về hình ảnh và siêu dữ liệu. Kết quả phát hiện tổng thể được kết hợp bằng các phương pháp như cộng (Sum), ghép nối (Concatenate) và chọn giá trị lớn nhất (Maximum). Kết quả thử nghiệm cho thấy, mô hình đạt độ chính xác 88% trong pha huấn luyện và 88,1% trong pha kiểm thử sử dụng dữ liệu kết hợp hình ảnh và bình luận về hình ảnh. Mô hình đạt kết quả tốt hơn khi sử dụng toàn bộ các loại dữ liệu, gồm văn bản, hình ảnh, bình luận về ảnh và siêu dữ liệu của các bài đăng trên mạng xã hội. Tuy vậy, tương tự như Uppada và cộng sự [4], phương pháp đề xuất chỉ được huấn luyện trên dữ liệu tiếng Anh, nên hiệu suất phát hiện có thể giảm khi áp dụng cho xử lý các bài viết trên các ngôn ngữ khác, như tiếng Việt.

Wu và cộng sự [6] đã phát triển một mô hình mạng kết hợp (Fusion Network) dựa trên cơ chế đồng chú ý (Co-Attention) đa phương thức để phát hiện tin giả. Mô hình này sử dụng mô hình BERT để xử lý văn bản và mô hình VGG19 để xử lý hình ảnh. Họ đã sử dụng dữ liệu từ Twitter và Weibo, kết hợp các đặc trưng từ văn bản, không gian và tần suất để phát hiện các bài đăng giả mạo. Ưu điểm của mô hình là cho hiệu suất phát hiện cao, đạt độ chính xác 80,9% trên tập dữ liệu Twitter và 89,9% trên tập dữ liệu Weibo. Tuy vậy, mô hình đề xuất chỉ được huấn luyện trên dữ liệu tiếng Anh và tiếng Trung, hạn chế khả năng áp dụng cho xử lý các bài viết trên các ngôn ngữ khác, như tiếng Việt. Ngoài ra, mô hình này không cung cấp chi tiết về phương pháp xử lý từng loại dữ liệu và cơ chế đồng chú ý, khiến việc cài đặt lại mô hình rất khó khăn.

Kiela và cộng sự [13] đã nghiên cứu phân loại tin tức đa phương thức trên qui mô lớn với tốc độ cao. Mô hình có khả năng xử lý kết hợp nhiều phương thức thể hiện của tin tức, như phương thức rời rạc với văn bản, và phương thức liên tục với hình ảnh được chuyển từ mạng nơ ron tích chập. Đặc biệt, mô hình tập trung vào các kịch bản cần phải phân loại lượng lớn dữ liệu một cách nhanh chóng. Các tác giả cũng nghiên cứu nhiều phương pháp khác nhau để thực hiện hợp nhất đa phương thức và phân tích sự đánh đổi của chúng về độ chính xác phân loại và hiệu quả tính toán. Các kết quả nghiên cứu chỉ ra rằng việc đưa thông tin liên tục vào sẽ cải thiện hiệu suất so với chỉ có văn bản trên một loạt các tác vụ phân loại đa phương thức, ngay cả với các phương pháp hợp nhất đơn giản. Nghiên cứu về phân loại tin tức đa phương thức đã mở ra một hướng đi đầy hứa hẹn cho giải quyết bài toán phát hiện tin bài giả mạo, độc hại trên không gian mạng, do các tin bài thường được đăng tải với nhiều hình thức thể hiện, như văn bản, ảnh, video, và sử dụng nhiều ngôn ngữ khác nhau.

Như vậy, có thể thấy hầu hết các mô hình phát hiện tin bài giả mạo, độc hại được phát triển cho tiếng Anh, nên khả năng áp dụng trực tiếp để xử lý tin bài tiếng Việt bị hạn chế. Hơn nữa, các tập dữ liệu huấn luyện gồm các tin bài tiếng Việt có số lượng đủ lớn để xây dựng mô hình phát hiện tin bài giả mạo, độc hại cũng không có sẵn, hoặc không được công khai. Ngoài ra, một số đề xuất cho hiệu suất phát hiện chưa cao, như [3], [5] và [6] chỉ đạt độ chính xác thấp hơn 90%. Trong bài báo này, chúng tôi đề xuất một mô hình nhằm nâng cao hiệu suất phát hiện tin bài phản động tiếng Việt dựa trên kết hợp mô hình PhoBERT và mô hình Swin Transformmer V2.

III. MÔ HÌNH PHÁT HIỆN TIN BÀI PHẢN ĐỘNG

A. Khái quát về một số mô hình học sâu

Mục này giới thiệu một số mô hình học sâu có liên quan và được sử dụng trong mô hình phát hiện tin bài phản động đề xuất. Các mô hình học sâu có liên quan và được sử dụng bao gồm Xception, Swin Transformer V2, BERT và PhoBERT.

1) Xception

Xception [24] là một kiến trúc mạng nơ-ron tích chập được phát triển từ kiến trúc Inception V3, chuyên dụng cho xử lý hình ảnh. Nó là một mô hình tiên tiến được huấn luyện trên một tập dữ liệu hình ảnh lớn, cho phép nhận biết và phân loại các đối tượng trong hình ảnh một cách hiệu quả. Xception cho hiệu suất vượt trội trong phân loại hình ảnh với độ chính xác cao trên nhiều bộ dữ liệu khác nhau. Nó cũng có khả năng tổng quát hóa tốt nhờ được huấn luyện trên một tập dữ liệu lớn và đa dạng. Xception có khả năng nhận biết và phân loại chính xác các đối tượng mới, chưa từng tồn tại trong bộ dữ liệu huấn luyện. Nhìn chung, kiến trúc của Xception được tối ưu hóa cho phép nó xử lý hình ảnh nhanh chóng và hiệu quả. Nhược điểm của Xception là kiến trúc phức tạp, đòi hỏi lượng tài nguyên tính toán lớn cho pha huấn luyện và pha phân loại. Ngoài ra, Xception cũng khó tùy chỉnh, nên việc điều chỉnh kiến trúc của Xception để phù hợp với các tác vụ cụ thể có thể phức tạp và tốn nhiều thời gian.

2) Swin Transformer V2

Swin Transformer V2 [9] là một biến thể cải tiến của mô hình Transformer, được thiết kế đặc biệt cho xử lý hình ảnh. Mô hình này kế thừa những ưu điểm của phiên bản trước, đồng thời khắc phục một số hạn chế về hiệu suất và khả năng mở rộng. Swin Transformer V2 đạt hiệu suất tốt hơn so với phiên bản trước trong việc phân loại hình ảnh, đặc biệt là trong việc xử lý các hình ảnh có độ phân giải cao. Swin Transformer V2 cũng có khả năng xử lý hiệu quả các hình ảnh có kích thước lớn, khả năng học được các đặc trưng phức tạp của hình ảnh một cách hiệu quả, giúp nó đạt được độ chính xác cao hơn trong phân loại. Nhược điểm của mô hình này là nó đòi hỏi nhiều tài nguyên tính toán hơn so với các mô hình xử lý hình ảnh truyền thống. Ngoài ra, việc tối ưu hóa để đạt được hiệu suất phân loại tối ưu có thể phức tạp và tốn thời gian.

3) BERT

BERT (Bidirectional Encoder Representations from Transformers) [25] là một mô hình ngôn ngữ lớn được huấn luyện để hiểu và xử lý văn bản bằng cách học ngữ cảnh của từng từ trong câu. BERT là một công cụ mạnh mẽ được sử dụng rộng rãi trong xử lý ngôn ngữ tự nhiên (NLP). Ưu điểm của BERT là có khả năng hiểu ngữ cảnh của các từ trong câu, giúp nó phân tích văn bản một cách chính xác hơn. BERT cũng cho hiệu suất cao trong nhiều nhiệm vụ NLP, như phân loại văn bản, trích xuất thông tin và dịch máy. BERT được huấn luyện trên một tập dữ liệu khổng lồ, cho phép nó xử lý các loại văn bản khác nhau một cách hiệu quả. Do BERT là mô hình ngôn ngữ lớn nên nó đòi hỏi nhiều tài nguyên tính toán cho huấn luyện và phân loại. Ngoài ra, việc điều chỉnh BERT cho các tác vụ cụ thể có thể phức tạp và tốn thời gian.

4) PhoBERT

PhoBERT [8] là một mô hình ngôn ngữ lớn dựa trên kiến trúc BERT, được huấn luyện sử dụng tập dữ liệu tiếng Việt, do Viện VinAI phát triển. PhoBERT được thiết kế để hiểu và xử lý văn bản tiếng Việt một cách hiệu quả, giúp cải thiện độ chính xác của các ứng dụng NLP cho tiếng Việt. Ưu điểm của PhoBERT là nó được huấn luyện trên một tập dữ liệu tiếng Việt khổng lồ và đa dạng, giúp nó hiểu ngữ cảnh và các sắc thái ngôn ngữ của tiếng Việt một cách chính xác. PhoBERT cũng cho hiệu suất cao trong nhiều nhiệm vụ NLP cho tiếng Việt như phân loại văn bản, trích xuất thông tin và dịch máy. Tương tự như BERT, PhoBERT là mô hình ngôn ngữ lớn nên nó đòi hỏi nhiều tài nguyên tính toán cho huấn luyện và phân loại. Ngoài ra, việc điều chỉnh PhoBERT cho các tác vụ cụ thể có thể phức tạp và tốn thời gian.

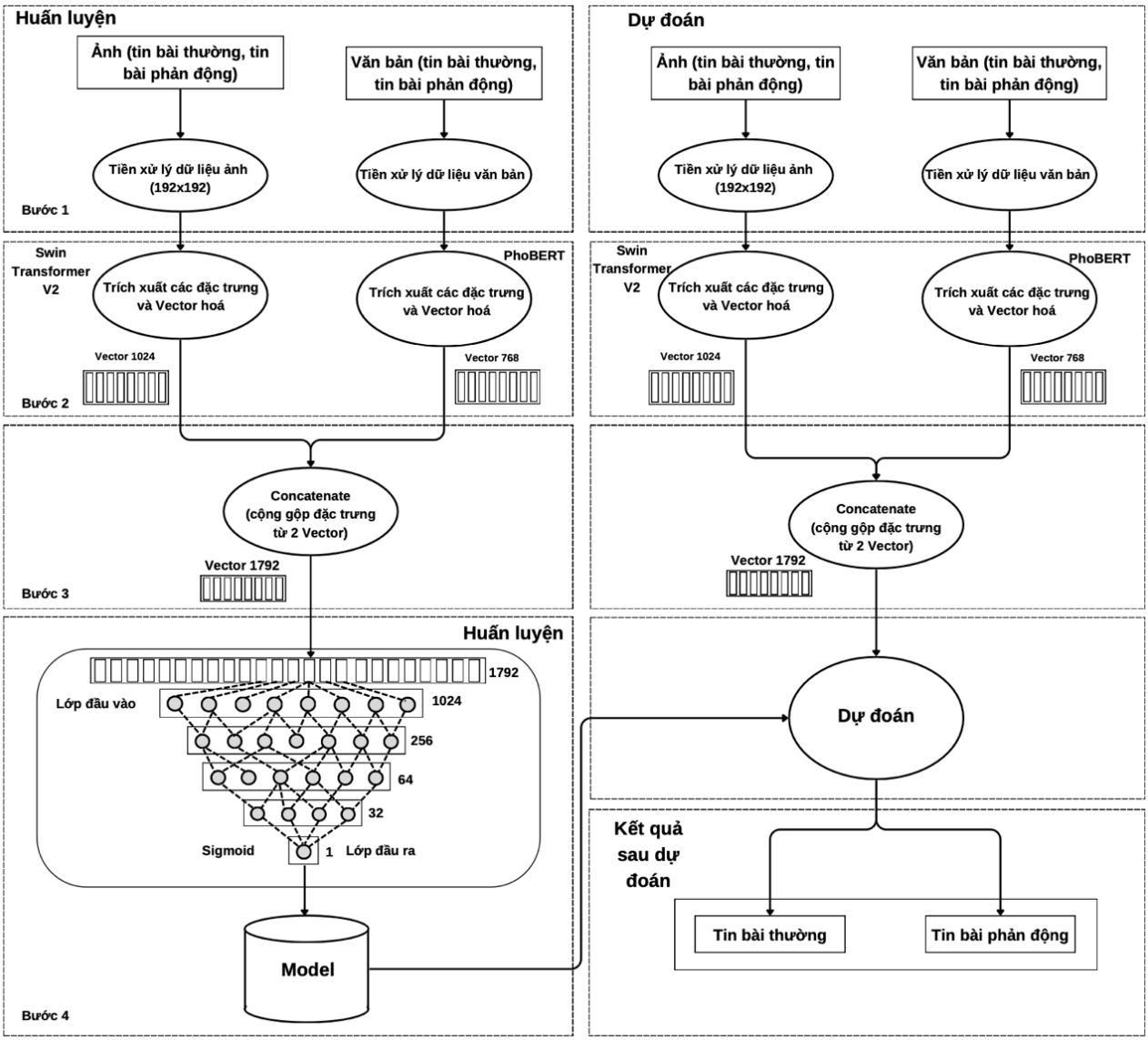
B. Mô hình phát hiện tin bài phản động đề xuất

1) Giới thiệu mô hình đề xuất

Kiến trúc của mô hình đa phương pháp đề xuất cho phát hiện tin bài phản động được mô tả trên Hình 1. Mô hình đề xuất sử dụng tập dữ liệu gồm các tin bài thông thường và các tin bài phản động. Mỗi loại tin bài lại gồm 2 dạng: các tin bài được thể hiện dưới dạng văn bản và các tin bài được thể hiện dưới dạng ảnh. Với dạng ảnh, nội dung văn bản của tin bài được tích hợp vào các bức ảnh. Mô hình đề xuất được triển khai theo 2 giai đoạn: giai đoạn huấn luyện và giai đoạn dự đoán. Trong giai đoạn huấn luyện, lưu đồ xử lý của mô hình được chia làm 2 nhánh: nhánh bên trái sử dụng mô hình Swin Transformer V2 để xử lý các tin bài dưới dạng ảnh và nhánh bên phải sử dụng mô hình PhoBERT để xử lý tin bài dưới dạng văn bản. Mô hình đề xuất sử dụng tất cả các lớp của các mô hình PhoBERT và Swin Transformer V2 đã được tinh chỉnh, ngoại trừ các lớp phân loại cuối cùng. Kết quả của 2 nhánh trên sẽ được kết hợp để cho ra một vector đặc trưng hợp nhất. Vector đặc trưng hợp nhất được đưa vào huấn luyện để sinh ra mô hình dự đoán. Trong giai đoạn dự đoán, tin bài dạng ảnh và dạng văn bản được tiền xử lý tương tự trong quá trình huấn luyện để sinh ra vector đặc trưng hợp nhất và vector này được phân loại sử dụng mô hình dự đoán để tạo ra kết quả là nhãn của tin bài thuộc dạng bình thường hay phản động.

2) Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng trong xử lý dữ liệu, gồm cả tin bài dưới dạng văn bản và tin bài dưới dạng hình ảnh. Với tin bài dưới dạng văn bản, đặc biệt khi làm việc với văn bản tự động thu thập từ Internet, nơi thường chứa nhiều ký tự không mong muốn và theo định dạng không đồng nhất. Các bước tiền xử lý đầu tiên bao gồm: chuẩn hóa văn bản như chuyển tất cả chữ cái về cùng một kiểu chữ, và loại bỏ các ký tự lạ. Tiếp theo, loại bỏ các từ viết tắt, dấu câu và liên kết (link) không cần thiết. Những ký tự này thường không cung cấp thông tin hữu ích cho phân tích văn bản và có thể gây nhiễu cho mô hình học máy. Bước kế tiếp quan trọng trong tiền xử lý là phân tách từ, tức là chia văn bản thành các từ riêng biệt để chuẩn bị cho các bước xử lý tiếp theo. Quá trình phân tách từ giúp tạo đầu vào phù hợp cho các mô hình học máy, làm cho chúng dễ dàng hơn trong việc phân tích và hiểu nội dung văn bản. Cuối cùng, dữ liệu văn bản được đệm để tạo ra các chuỗi văn bản có độ dài cố định làm đầu vào cho quá trình huấn luyện. Với tin bài dưới dạng hình ảnh, các hình ảnh được thay đổi kích thước thành 192x192 điểm ảnh trước khi đưa vào quá trình huấn luyện.



Hình 1. Kiến trúc của mô hình đa phương pháp đề xuất

3) Quá trình huấn luyện

Quá trình huấn luyện mô hình kết hợp đa phương pháp đề xuất, như biểu diễn trên Hình 1, gồm các bước sau:

- Bước 1: Các tin bài được tiền xử lý như mô tả tại Mục III.B.2.

- Bước 2: Các tin bài dưới dạng văn bản đã qua tiền xử lý được đưa vào mô hình PhoBERT để xử lý và nhận các vector có kích thước 768. Các tin bài dưới dạng hình ảnh đã qua tiền xử lý được đưa vào mô hình Swin Transformer V2 để xử lý và nhận được vector đầu ra là 1024.

- Bước 3: Thực hiện kết hợp vector đầu ra của PhoBERT và vector đầu ra của Swin Transformer V2 bằng phương pháp Concatenate, kết quả là 1 vector tổng hợp có kích thước 1792. Phương pháp Concatenate giữ nguyên toàn bộ thông tin của hai vector đầu ra từ PhoBERT (768) và Swin Transformer V2 (1024), tạo thành một vector hợp nhất (1792). So với các phương pháp khác như tính trung bình hay tích vô hướng, Concatenate cho phép hệ thống lưu giữ đặc trưng đầy đủ từ cả văn bản và hình ảnh mà không làm mất thông tin chi tiết. Đây là phương pháp đơn giản nhưng hiệu quả, giúp giảm thiểu độ phức tạp trong huấn luyện. Tuy vậy, kích thước vector hợp nhất lớn (1792) có thể làm tăng chi phí tính toán. Mặc dù vậy, với phần xử lý Dense Layer sau đó, vấn đề này có thể được giải quyết hiệu quả.

- Bước 4: Vector đặc trưng hợp nhất (1792) được đưa vào huấn luyện để sinh ra mô hình dự đoán (Model). Mô hình dự đoán được sử dụng để phân loại tin bài cần xử lý trong giai đoạn dự đoán.

Trong quá trình huấn luyện, trọng số của các lớp trong cả hai nhánh PhoBERT và Swin Transformer V2 đã được khóa. Điều này có nghĩa là trọng số học được trong giai đoạn huấn luyện trước đó của các mô hình này sẽ không thay đổi trong quá trình huấn luyện mô hình đa phương pháp. Thuật toán tối ưu hóa Adam với hàm mất mát binary cross-entropy được sử dụng và batch size huấn luyện là 32.

Việc sử dụng các mô hình PhoBERT và Swin Transformer V2 trong mô hình kế hợp đa phương pháp mang lại nhiều lợi ích so với việc sử dụng từng mô hình riêng lẻ. Mô hình kế hợp đa phương pháp có thể phân loại hiệu quả các tin bài phản động / tin bài thường nhờ khả

năng kết hợp các ưu điểm của các mô hình riêng lẻ, đặc biệt trong các trường hợp dữ liệu phức tạp bao gồm cả các tin bài được thể hiện dưới dạng văn bản và dạng hình ảnh. Cụ thể, đóng góp của mỗi mô hình thành phần trong mô hình đề xuất như sau:

- Mô hình PhoBERT được tối ưu hóa cho dữ liệu văn bản tiếng Việt, thực hiện xử lý và mã hóa thông tin văn bản thành vector có kích thước 768. PhoBERT rất mạnh trong việc biểu diễn ngữ nghĩa và cú pháp của văn bản nhờ kiến trúc transformer tiên tiến, tối ưu hóa dựa trên mô hình BERT. Trong mô hình đề xuất, PhoBERT đóng vai trò chính trong việc phân tích nội dung văn bản, góp phần trực tiếp vào việc phân loại dựa trên đặc điểm ngữ nghĩa.

- Mô hình Swin Transformer V2 là một kiến trúc xử lý hình ảnh hiệu quả nhờ khả năng mã hóa hình ảnh qua các vùng không gian và mức độ phân giải khác nhau. Vector 1024 đầu ra chứa thông tin trực quan quan trọng, như các đặc điểm hình học hoặc mô hình màu sắc, giúp phát hiện các đặc điểm đặc trưng của hình ảnh, chẳng hạn các biểu tượng, hình minh họa liên quan đến nội dung phản động trong tin bài.

4) Quá trình dự đoán

Trong giai đoạn dự đoán, việc xử lý tin bài dạng ảnh và dạng văn bản được tiền xử lý tương tự trong quá trình huấn luyện tại các bước từ bước 1 đến bước 3 để sinh ra vector đặc trưng hợp nhất có kích thước 1792. Trong bước 4, vector đặc trưng hợp nhất được phân loại sử dụng mô hình dự đoán (Model) để tạo ra kết quả là nhãn của tin bài thuộc dạng bình thường hay phản động.

5) Các độ đo đánh giá

Để đánh giá được hiệu suất mô hình phân loại, các độ đo được sử dụng bao gồm: Accuracy, Presicion, Recall và F1-score. Độ đo Accuracy là độ chính xác tổng thể, được tính theo công thức:

$$Accuracy = \frac{\text{TP+TN}}{\text{tổng số mẫu}} \quad (1)$$

Độ đo Presicion là độ chính xác, được tính theo công thức:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Độ đo Recall là độ nhạy hay độ bao phủ, được tính theo công thức:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score là độ đo biểu diễn trung bình điều hòa giữa 2 độ đo Precision và Recall, được tính theo công thức:

$$\text{F1} = \frac{2(precision*recall)}{precision+recall} \quad (4)$$

Các tham số TP, FP, FN và TN được biểu diễn trong ma trận nhầm lẫn trên Bảng I.

Bảng I. Các tham số TP, FP, FN và TN trong ma trận nhầm lẫn

		Lớp thực tế	
		Phản động	*Bình thường*
Lớp dự đoán	*Phản động*	TP (True Possitive)	FP (False Possitive)
	Bình thường	FN (False Negative)	TN (True Negative)

IV. THỬ NGHIỆM VÀ KẾT QUẢ

A. Thu thập bộ dữ liệu thử nghiệm

Bộ dữ liệu thu thập gồm 4000 tin bài tiếng Việt sử dụng để huấn luyện và kiểm thử mô hình đa phương pháp đề xuất cho phát hiện tin bài phản động, bao gồm:

- Các tin bài có nội dung phản động: gồm 1000 bài viết dưới dạng văn bản và 1000 bài dưới dạng ảnh từ nhiều nguồn khác nhau, như trên các diễn đàn hải ngoại, các fanpage của các mạng xã hội. Các tin bài thuộc nhóm này được dán nhãn 'Phản động';

- Các tin bài có nội dung bình thường: 1000 bài viết dưới dạng văn bản và 1000 bài dưới dạng ảnh thu thập từ các trang tin tức và các diễn đàn. Các tin bài thuộc nhóm này được dán nhãn 'Bình thường'.

Tập dữ liệu thu thập được chia ngẫu nhiên thành 2 phần: 80% dữ liệu cho huấn luyện và 20% dữ liệu cho kiểm thử.

B. Kết quả thử nghiệm và nhận xét

1) Các kịch bản và kết quả thử nghiệm

Tập dữ liệu thu thập được như mô tả tại Mục 4.1 với 80% dữ liệu cho huấn luyện và 20% dữ liệu cho kiểm thử được sử dụng để huấn luyện và kiểm thử hiệu suất phát hiện của các mô hình riêng lẻ và mô hình kết hợp đề xuất, bao gồm:

- Lần lượt sử dụng các mô hình Xception và Swin Transformer V2 để xử lý tập dữ liệu gồm 1000 tin bài bình thường và 1000 tin bài phản động dưới dạng ảnh;

- Lần lượt sử dụng các mô hình BERT và PhoBERT để xử lý tập dữ liệu gồm 1000 tin bài bình thường và 1000 tin bài phản động dưới dạng văn bản;

- Sử dụng mô hình kết hợp đề xuất để xử lý tập dữ liệu gồm 1000 tin bài bình thường và 1000 tin bài phản động dưới dạng văn bản, và 1000 tin bài bình thường và 1000 tin bài phản động dưới dạng ảnh.

Bảng II cung cấp kết quả thử nghiệm theo các kịch bản nên trên.

2) Nhận xét

Từ các kết quả thử nghiệm cho trên Bảng 2, có thể rút ra một số nhận xét:

- Các mô hình BERT và PhoBERT cho hiệu suất phát hiện tốt hơn đáng kể so với các mô hình Xception và Swin Transformer V2. Điều này là do các tin bài dưới dạng văn bản thường có ít nhiễu, chứa nhiều thông tin nên có thể được xử lý hiệu quả bởi các mô hình được huấn luyện trước như BERT và PhoBERT. Ngoài ra, cũng có thể thấy, PhoBERT cho hiệu suất phát hiện tốt hơn đáng kể so với BERT do PhoBERT được huấn luyện trước trên tập dữ liệu tiếng Việt lớn, nên có khả năng xử lý tin bài tiếng Việt hiệu quả hơn.

- Mặc dù Xception và Swin Transformer V2 có hiệu suất phát hiện thấp hơn BERT và PhoBERT, nhưng các độ đo cũng ở mức khá tốt. Điều này là do các tin bài dưới dạng ảnh thường có nhiều nhiễu, chứa ít thông tin hơn so với văn bản. Ngoài ra, cũng có thể thấy, Swin Transformer V2 cho hiệu suất phát hiện tốt hơn đáng kể so với Xception do mô hình Swin Transformer V2 cũng có khả năng xử lý dữ liệu văn bản nhúng trong ảnh tốt hơn so với mô hình Xception.

- Mô hình kết hợp đa phương pháp đề xuất dựa trên các mô hình PhoBERT và Swin Transformer V2 sử dụng các đặc trưng văn bản và ảnh cho hiệu suất phát hiện vượt trội so với các mô hình riêng lẻ. Cụ thể, các độ đo của mô hình kết hợp đề xuất: Accuracy đạt 97%, Precision đạt 97%, Recall đạt 97.5% và F1-score đạt 97%.

Bảng II. Kết quả thử nghiệm hiệu suất phát hiện của mô hình đề xuất và các mô hình riêng lẻ

Mô hình	Đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Xception	Ảnh	82.25	83.68	79.90	81.75
Swin Transformer V2	Ảnh	85.00	84.24	85.93	85.07
BERT	Văn bản	91.25	91.26	91.23	91.24
PhoBERT	Văn bản	96.50	96.49	96.63	96.50
PhoBERT+ Swin Transformer V2	**Văn bản+Ảnh**	**97.00**	**97.00**	**97.5**	**97.00**

Bảng III. So sánh hiệu suất phát hiện của mô hình đề xuất với nghiên cứu có liên quan

Mô hình	Đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BERT+Xception (Concatenate) [4]	Văn bản+Ảnh	91.70	93.39	93.29	93.25
BERT+Xception (Maximum) [4]	Văn bản+Ảnh	91.68	93.76	92.83	93.29
PhoBERT+ Swin Transformer V2	**Văn bản+Ảnh**	**97.00**	**97.00**	**97.5**	**97.00**

Bảng III cung cấp so sánh hiệu suất phát hiện của mô hình đề xuất với nghiên cứu có liên quan [4]. Có thể thấy, mô hình kết hợp đề xuất có hiệu suất phát hiện cao hơn đáng kể so với 2 phương án của các mô hình kết hợp đề xuất trong [4], bao gồm BERT+Xception (Concatenate) và BERT+Xception (Maximum). Cụ thể, độ đo F1 của mô hình kết hợp đề xuất dựa trên PhoBERT+Swin Transformer V2 và các mô hình dựa trên BERT+Xception (Concatenate), BERT+Xception (Maximum) [4] lần lượt là 97% và 93.25, 93.29%.

Hiệu suất phát hiện của mô hình đa phương pháp đề xuất tốt hơn với so với các mô nghiên cứu liên quan bởi hai lý do sau: (i) mô hình PhoBERT có khả năng xử lý dữ liệu tiếng Việt tốt hơn hẳn so với mô hình BERT, và mô hình Swin Transformer V2 cũng có khả năng xử lý dữ liệu văn bản nhúng trong ảnh tốt hơn so với mô hình Xception, nên việc lựa chọn sử dụng PhoBERT và Swin Transformer V2 trong mô hình kết hợp là phù hợp, và (ii) việc lựa chọn kết hợp mô hình PhoBERT kết hợp với mô hình Swin Transformer V2 khai thác được điểm mạnh của cả 2 mô hình riêng lẻ, giúp tăng khả năng nhận diện các đặc trưng cả trên văn bản và trên ảnh, nhờ vậy có khả năng phân biệt tốt hơn giữa tin bài phản động và tin bình thường.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo này đề xuất mô hình đa phương pháp dựa trên sự kết hợp của mô hình PhoBERT và mô hình Swin Transformer V2 sử dụng đặc trưng trích xuất từ các tin bài dưới dạng văn bản và các tin bài dưới dạng hình ảnh. Các kết quả thử nghiệm trên bộ dữ liệu tin bài tiếng Việt, gồm 1000 tin bài phản động dưới dạng văn bản, 1000 tin bài phản động dưới dạng ảnh, 1000 tin bài thường dưới dạng văn bản và 1000 tin bài thường dưới dạng ảnh, khẳng định mô hình kết hợp đề xuất cho các độ đo phát hiện vượt trội so với các mô hình riêng lẻ và các mô hình đã có.

Trong tương lai, chúng tôi tiếp tục nghiên cứu, cải tiến mô hình đề xuất, kết hợp thêm các đặc trưng khác trong phân tích, phát hiện tin bài phản động, nhằm (i) tiếp tục nâng cao độ chính xác và giảm tỷ lệ nhận diện sai, và (ii) giảm yêu cầu sử dụng tài nguyên tính toán trong huấn luyện và đặc biệt trong khâu phát hiện tin phản động để cải thiện khả năng ứng dụng trên thực tế.

TÀI LIỆU THAM KHẢO

[1] Báo điện tử đảng cộng sản Việt Nam. [Online] https://dangcongsan.vn/bao-ve-nen-tang-tu-tuong-cua-dang/bao-dien-tu-dau-tranh-phan-bac-nhung-luan-dieu-sai-trai-phan-dong-tren-mang-xa-hoi-hien-nay-596988.html Ngày đăng bài 15/11/2021.

[2] J. Cao , P. Qi ,Q. Sheng, T. Yang, J. Guo, and J. Li, Exploring the Role of Visual Content in Fake News Detection, 2020, https://arxiv.org/pdf/2003.05096.

[3] H. Nguyen, S. Gokhale, An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests, November 25, 2022, https://peerj.com/articles/cs-1127/.

[4] S.K. Uppada, P. Patel and S. B, An image and text-based multimodal model for detecting fake news in OSN's, 2022, ISSN:0925-9902 E-ISSN:1573-7675, https://link. springer.com/article/10.1007/s10844-022-00764-y.

[5] K. Armin, S. Djordje, Z. Matthias, Multimodal Detection of Information Disorder from Social Media, 31 May 2021, https://arxiv.org/pdf/2105.15165.

[6] Y. Wu, P. Zhan, Y.unjian Zhang, L. Wang, Z. Xu, Multimodal Fusion with Co-Attention Networks for Fake News Detection, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2560–2569, https://aclanthology.org/2021.findings-acl.226.pdf.

[7] K. Nakamura, S. Levy, W.Y. Wang, Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection, 2020, https://arxiv.org/pdf/1911.03854.

[8] D.Q. Nguyen, A.T. Nguyen, PhoBERT: Pre-trained language models for Vietnamese, 2020, https://arxiv.org/abs/2003.00744.

[9] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin Transformer V2: Scaling Up Capacity and Resolution, 11 Apr 2022, https://arxiv.org/pdf/2111.09883.

[10] F. Chollet - Google Inc, Xception: Deep Learning with Depth wise Separable Convolutions. https://arxiv.org/abs/1610.02357 , 2017.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova - Google AI Language, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 24 May 2019, https://arxiv.org/pdf/1810.04805.

[12] J. Jing, H. Wu , J. Sun , X. Fang, H. Zhang, Multimodal fake news detection via progressive fusion networks, 2023, https://www.sciencedirect.com/science/article/pii/S0306457322002217#b22.

[13] D. Kiela, E. Grave, A. Joulin and T. Mikolov, Efficient Large-Scale Multi-Modal Classification, 2018, https://arxiv.org/pdf/1802.02892.

[14] J. Ahmed and M. Ahmed, Online News Classification Using Machine Learning Techniques, IIUMEJ, vol. 22, no. 2, pp. 210– 225, Jul. 2021.

[15] A. Aashish et al., Good , Neutral or Bad - News Classification, NewsIR@SIGIR (2019).

[16] W. Antoun, F. Baly, and H. J. a. p. a. Hajj, "Arabert: Transformer-based model for arabic language understanding," 2020.

[17] M. Abdul-Mageed, A. Elmadany, and E. M. B. J. a. p. a. Nagoudi, "ARBERT & MARBERT: deep bidirectional transformers for Arabic," 2020.

[18] X. Li, L. Bing, W. Zhang, and W. Lam, “Exploiting BERT for End-to-End Aspect-based Sentiment Analysis,” ArXiv, abs/1910.00883, 2019.

[19] K. Nugroho, A. Sukmadewa, and N. Yudistira, "Large-scale News Classification Using Bert Language Model: Spark NLP Approach," Arxiv, https://doi.org/10.1145/3479645.3479658, 2021.

[20] A. Ali, SAM. Noah, LQ. Zakaria, “A BERT- Based Model: Improving Crime News Documents Classification through Adopting Pre- trained Language Models,” Research Square, doi:10.21203/rs.3.rs-2582775/v1, 2023

[21] B. Juarto and Yulianto, “Indonesian News Classification Using IndoBERT,” International Journal of Intelligent Systems and Applications in Engineering, Vol 1, No 2, 2023.

[22] J. Devlin, M.W. Chang, K. Lee and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.

[23] Y. Liu, et al., “Roberta: A robustly optimized bert pretraining approach,” arXiv 2019, arXiv preprint arXiv: 1907.11692.

[24] G. Boesch, Xception Model: Analyzing Depthwise Separable Convolutions, https://viso.ai/deep-learning/xception-model/.

[25] A. Tam, A Brief Introduction to BERT, https://machinelearningmastery.com/a-brief-introduction-to-bert/.

A MULTIMODAL MODEL FOR DETECTING VIETNAMESE REACTIONARY NEWS ARTICLES

Abstract: Fake news, news articles with toxic, reactionary content are currently posted and spreaded very strongly due to the popularity of the Internet and especially the explosion of social networks and online services in cyberspace. News articles with toxic content, and especially reactionary news articles aimed at Vietnam, such as news articles spreading false information, slandering leaders, inciting destruction of the great national unity bloc, have a great impact on social life because they can spread quickly and have many forms of expression, such as news articles in the form of text, photos, or a combination. Due to the seriousness of articles posting fake news, or articles with toxic, reactionary content in cyberspace, there have been a number of studies in Vietnam and in the world for detection and prevention of these articles. However, most of the proposals focus on handling fake, toxic news articles posted using the English language. Furthermore, due to the large number of news articles posted in the form of images, or text embedded in images and videos, the handling is difficult, leading to a relatively low correct detection rate. This paper proposes a multimodal model based on the combination of PhoBERT and Swin Transformer V2 models for detecting reactionary news articles in the form of text and images. Experimental results show that the proposed combined model using text and image features gives superior detection performance compared to individual models and existing models, with accuracy at 97%, precision at 97%, recall at 97.5% and F1-score at 97%.

Keywords: Detecting reactionary news, multimodal model, PhoBERT, Swin Transformer V2, Swin Transformer v2 + PhoBERT

Lê Ngọc An tốt nghiệp Thạc sỹ chuyên ngành Hệ thống thông tin tại Học viện Công nghệ Bưu chính Viễn thông năm 2021. Hiện là NCS tại Học viện Công nghệ Bưu chính Viễn thông và công tác tại Khoa Công nghệ thông tin, Trường Đại học Mở Hà Nội. Lĩnh vực nghiên cứu: An ninh mạng, an toàn thông tin trên không gian mạng, Xử lý ngôn ngữ tự nhiên, Thị giác máy tính

Email: anln@hou.edu.vn

Hoàng Xuân Dậu nhận học vị Tiến sỹ năm 2006, học hàm Phó giáo sư năm 2022. Hiện công tác tại Khoa An toàn thông tin, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: phát hiện tấn công, xâm nhập, phát hiện mã độc, bảo mật hệ thống và phần mềm, bảo mật web và các ứng dụng dựa trên học máy cho an toàn thông tin.

Email: dauhx@ptit.edu.vn

Dương Trần Đức tốt nghiệp Thạc sỹ chuyên ngành Hệ thống thông tin tại Đại học Tổng hợp Leeds, Vương Quốc Anh năm 2004, và Tiến sỹ chuyên ngành Kỹ thuật máy tính tại Học viện Công nghệ Bưu chính Viễn thông năm 2018. Hiện công tác tại Khoa Công nghệ Thông tin 1, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Xử lý ngôn ngữ tự nhiên, Thị giác máy tính

Email: ducdt@ptit.edu.vn

Đinh Tuấn Long nhận học vị Tiến sĩ năm 2014. Hiện công tác tại Khoa Công nghệ thông tin, Trường Đại học Mở Hà Nội. Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, Ứng dụng CNTT trong giáo dục, An toàn bảo mật dữ liệu.

Email: dinhtuanlong@hou.edu.vn