PHÂN LOẠI TÍNH CHẤT VỤ VIỆC BÀI BÁO MẠNG DỰA TRÊN MÔ HÌNH PhoBERT

Lê Ngọc An[*], Nguyễn Đình Toàn[*], Lê Trường Thiên[*], Dương Trần Đức[+]
[*] Viện Tài nguyên Môi trường và Công nghệ thông tin INRES.AI
[+] Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Quản lý thông tin mạng là một vấn đề cấp thiết hiện nay khi các nội dung truyền thông ngày càng được số hóa và phổ biến. Nhờ tính chất thuận tiện của truyền thông, báo chí số, thông tin về các vụ việc được truyền tải một cách nhanh chóng và người đọc có thể dễ dàng tiếp cận. Cùng với sự phát triển này, các cơ quan quản lý cũng phải có khả năng nhanh chóng nắm bắt được thông tin để có các xử lý kịp thời. Trong đó, việc nhanh chóng thu thập và xác định tính chất vụ việc đang xảy ra trên truyền thông là một công việc quan trọng đối với các cơ quan quản lý thông tin. Bài báo này đề xuất một hệ thống và phương pháp tự động thu thập, đánh giá phân loại tính chất vụ việc qua nội dung bài báo mạng sử dụng mô hình dựa trên PhoBERT, trong đó sử dụng thêm một lớp phân loại tuyến tính và tinh chỉnh tập dữ liệu gồm hơn 6.000 bài báo được thu thập tự động và hỗ trợ gán nhãn bởi các chuyên viên trong lĩnh vực quản lý thông tin mạng. Kết quả phân loại cao nhất theo độ đo F1 đến 93.1% theo tính chất vụ việc với ba nhãn tích cực, tiêu cực, bình thường cho thấy phương pháp là khả thi và có thể áp dụng vào thực tế.

Từ khóa: phân loại bài báo mạng, quản lý thông tin mạng, mô hình bert.

I. MỞ ĐẦU

Trong bối cảnh các kênh trao đổi thông tin trực tuyến ngày càng không ngừng gia tăng, việc quản lý thông tin mạng là một vấn đề trở nên quan trọng và cấp thiết. Trong các vấn đề về quản lý thông tin mạng, việc nhanh chóng nắm bắt được các vụ việc, bao gồm các vụ việc có tính chất tích cực và tiêu cực, giúp cơ quan quản lý có thể nhanh chóng có phương án ứng phó và xử lý kịp thời.

Để thực hiện công việc này, các cơ quản quản lý phải thu thập các bài viết, đánh giá, phân loại, tổng hợp để phục vụ báo cáo và xử lý thông tin. Do sự bùng nổ của các kênh trao đổi thông tin như hiện nay, thực hiện việc này theo cách thủ công là vô cùng thiếu hiệu quả và tốn kém về nhân lực. Với sự phát triển của các mô hình xử lý ngôn ngữ tự nhiên, việc thu thập, đánh giá, phân loại các bài viết theo tính chất là một tác vụ hoàn toàn khả thi và có thể giải quyết được các vấn đề nêu trên.

Về các kỹ thuật phân loại văn bản, hiện nay các nghiên cứu chủ yếu khai khác kỹ thuật học máy, trong đó các mô hình dựa trên BERT là kỹ thuật mới nhất và nhận được sự quan tâm lớn của các nhà nghiên cứu. BERT là một mô hình xử lý ngôn ngữ tự nhiên được giới thiệu bởi Google vào năm 2018. BERT học ra các véc tơ đại diện theo ngữ cảnh 2 chiều của từ và tỏ ra hiệu quả vượt trội so với các mô hình trước đây như Word2Vec, Glove, v.v. BERT đã thành công trong việc cải thiện những công việc gần đây trong việc tìm ra đại diện của từ trong không gian số (không gian mà máy tính có thể hiểu được) thông qua ngữ cảnh của nó.

Bài báo này trình bày phương pháp phân loại bài viết trên các báo điện tử theo tính chất vụ việc được phản ánh trong bài báo theo hai thể loại là tích cực và tiêu cực. Tập dữ liệu được thu thập từ các báo điện tử phổ biến ở Việt Nam như VNExpress, Vietnamnet, Dân trí, và hơn 10 tờ báo điện tử khác. Tập dữ liệu được gán nhãn thủ công với ba nhãn là tích cực, tiêu cực, và bình thường. Tổng cộng hơn 6.000 bài báo trong các lĩnh vực Kinh tế, Xã hội, Giáo dục, Y tế, Xây dựng, Quản lý đô thị được sử dụng để huấn luyện và kiểm thử mô hình. Kết quả thực nghiệm cho độ chính xác phân loại từ 90.5% đến 94.2% với các độ đo và loại mô hình khác nhau.

Các đóng góp chính của nghiên cứu bao gồm:

- Đề xuất phương pháp phân loại bài báo mạng theo tính chất tích cực và tiêu cực của vụ việc sử dụng mô hình dựa trên PhoBERT cho tiếng Việt. Các thực nghiệm được thực hiện trên tập dữ liệu được thu thập từ các báo mạng và gán nhãn thủ công có sự hỗ trợ của các chuyên viên quản lý thông tin mạng.
- So sánh và đánh giá hiệu quả của mô hình dựa trên PhoBERT với mô hình dựa trên BERT đa ngôn ngữ và kỹ thuật sử dụng đặc trưng nhúng từ (word embeddings) truyền thống.

Bài báo có cấu trúc như sau. Phần II trình bày về các 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. TỔNG QUAN

Phân loại văn bản dựa trên học máy là một hướng nghiên cứu phổ biến hiện nay, với các ứng dụng điển hình như phân loại cảm xúc (sentiment analysis), phân loại email, lọc thư rác v.v. Trong đó, các nhà nghiên cứu trong nước chủ yếu tập trung vào bài toán phân loại cảm xúc. Thời kỳ đầu các nhà nghiên cứu sử dụng các phương pháp dựa trên luật [1]. Tuy nhiên, trong thời gian gần đây, phương pháp dựa trên học máy tỏ ra vượt trội hơn nhờ khả năng khai thác được các mối quan hệ ngữ cảnh trong văn bản. Nghiên cứu của Duyen et al. [2] ứng dụng các thuật toán Nai Bayes, Max Entropy Model, và SVM để phân loại các bài đánh giá trên hệ thống đặt phòng Agoda, trong đó SVM đạt kết quả tốt nhất. Nghiên cứu của Quan

Tác giả liên hệ: Dương Trần Đức,
Email: ducdt@ptit.edu.vn
Đến tòa soạn: 27/6/2023, chỉnh sửa: 20/8/2023, chấp nhận đăng: 06/9/2023.

et al. [3] sử dụng thuật toán học sâu, kết hợp của LSTM và CNN để phân loại các đánh giá trên các trang thương mại điện tử của Việt nam. Thai et al. [4] đề xuất phương pháp sử dụng mô hình BERT để phân loại bài đánh giá trên tập dữ liệu thu thập từ trang Foody.vn và các trang thương mại điện tử khác của Việt nam với độ chính xác khi sử dụng BERT cho kết quả cao hơn các phương pháp khác (SVM, FastText, Glove). Nhìn chung, các nghiên cứu trên tiếng Việt còn hạn chế về thể loại văn bản và bài toán ứng dụng.

Đối với các ngôn ngữ khác, đặc biệt là với ngôn ngữ phổ biến như tiếng Anh, các nghiên cứu về phân loại văn bản được thực hiện ở phạm vi rộng hơn. Cũng như các nghiên cứu trên tiếng Việt, các nghiên cứu trên ngôn ngữ khác gần đây tập trung sử dụng các thuật toán học sâu để khai thác thế mạnh của các mô hình giàu ngữ cảnh.

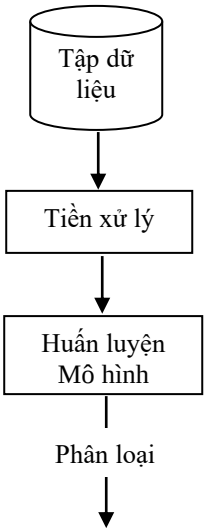
She et al. [5] sử dụng kỹ thuật kết hợp CNN-LSTM để phân loại các bản tin tiếng Trung Quốc cho hiệu quả cao. Cai et al. [6] thực hiện nghiên cứu phân loại tin tức bằng phương pháp kết hợp các kiến trúc học sâu như R-CNN, CNN, RNN. Lenc et al. [7] đề xuất phương pháp sử dụng CNN kết hợp MLP để trích xuất các đặc trưng từ các bài báo tiếng Séc. Kết quả có độ chính xác 84%. Koswari et al. [8] đề xuất một phương pháp học kết hợp sử dụng các mô hình học sâu cho phân loại các bài báo và đạt kết quả có độ chính xác 87%. Nghiên cứu của Ahmed et al. [9] thực hiện phân loại tin tức mạng theo chủ đề bằng các thuật toán học máy truyền thống như Naïve Bayes, K-Nearest Neighbour, Support Vector Machine với kết quả cao nhất 93% của Naïve Bayes. Aashish et al. [10] thực hiện nghiên cứu phân loại bài báo tiếng Anh với ba nhãn tốt, xấu, trung tính cũng với các thuật toán học máy truyền thống và Naïve Bayes cũng cho kết quả tốt nhất với 82.9%.

Antoun et al. [11] phát triển mô hình dựa trên BERT đề phân loại văn bản tiếng A rập gọi là Arabert. Mô hình được huấn luyện trên 24 Giagabytes dữ liệu có độ chính xác 96.2%. Tương tự, trong một nghiên cứu khác, Abdul Mageed et al. [12] huấn luyện mô kiến trúc dựa trên BERT có tên là MARTBERT trên tập dữ liệu 1B các bài viết mạng xã hội Twitter. Li et al. [13] nghiên cứu việc ứng dụng kiến trúc BERT cho bài toán phân loại cảm xúc dựa trên khía cạnh đem lại kết quả vượt trội so với các cạnh tiếp cận trước đây. Nugroho et al. [14] đề xuất phương pháp sử dụng mô hình BERT để phân loại bài báo mạng tiếng Anh từ tập dữ liệu AG[1]. Kết quả phân loại tốt nhất với mô hình BERT-Base cho đô chính xác 92.53%. A. Ali et al. [15] sử dụng BERT để phân loại tin tức tội phạm trên tập dữ liệu tiếng Malaysia với độ chính xác lên tới 99% (97% với độ đo F1). B. Juarto et al. [16] đề xuất phương pháp phân loại bài báo tiếng Indonesia bằng mô hình IndoBERT. Tổng số mẫu được sử dụng là hơn 8.000 bài báo (70% dùng để huấn luyện, 30 kiểm thử), cho kết quả tốt nhất với mô hình IndoBERT là 95%.

Các khảo sát trên cho thấy hiện nay BERT là một mô hình có tiềm năng lớn trong lĩnh vực xử lý ngôn ngữ tự nhiên nói chung và phân loại văn bản nói riêng. Nghiên cứu này đề xuất phương pháp sử dụng mô hình dựa trên BERT để phân loại bài báo tiếng Việt nhằm đánh giá tính chất tích cực và tiêu cực của nó, nhằm phục vụ công tác quản lý thông tin trên không gian mạng. Theo tìm hiểu của chúng tôi, chưa có nghiên cứu nào về chủ đề này được thực hiện trước đây.

III. PHƯƠNG PHÁP

Phần này sẽ trình bày chi tiết về phương pháp thực hiện, bao gồm các bước trong kiến trúc của mô hình. Hình 1 minh họa quá trình xây dựng mô hình. Các bước được giải thích chi tiết hơn trong các phần tiếp theo.



Hình 1. *Quá trình xây dựng mô hình*

A. Tiền xử lý dữ liệu

Tiền xử lý là một bước quan trọng trong quá trình xử lý văn bản, đặc biệt là với các loại văn bản tự động thu thập từ mạng Internet, vốn có thể chứa nhiều các ký tự và định dạng không mong muốn.

Một số bước tiền xử lý ban đầu có thể được thực hiện như chuẩn hóa câu, lọc bỏ các ký tự lạ, từ viết tắt, dấu câu, các liên kết (links). Bước cuối cùng trong quá trình tiền xử lý là hoạt động phân tách từ (word segmentation) nhằm tạo đầu vào cho các bước xử lý tiếp theo.

B. Xây dựng mô hình

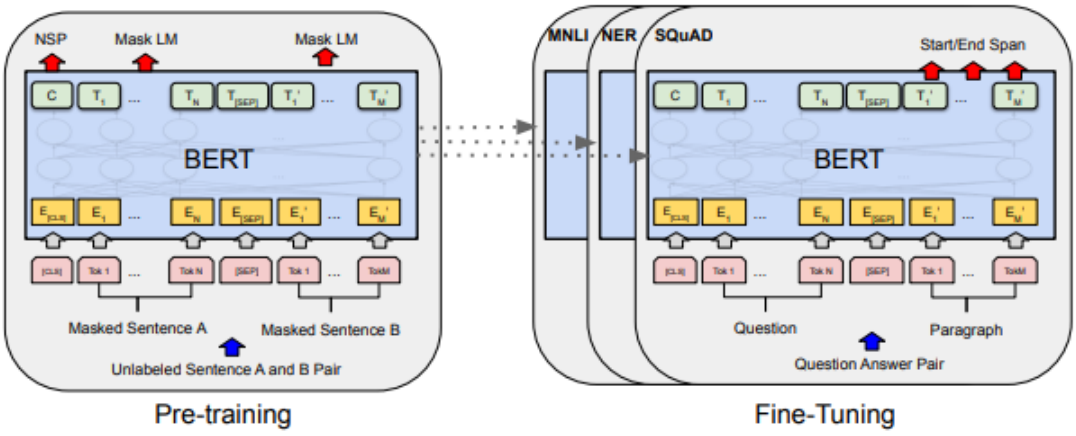
Mô hình phân loại trong nghiên cứu được xây dựng dựa trên PhoBERT, một mô hình được phát triển dựa trên kiến trúc BERT. Các phần tiếp theo sẽ trình bày về PhoBERT và các kiến trúc liên quan.

1) BERT

BERT (Bidirectional Encoder Representations from Transfomer) là một mô hình biểu diễn từ được phát triển sử dụng kỹ thuật Transformer bằng cách tạo các lớp mã hóa (transfomer encoder) và xếp chồng chúng với nhau để tạo thành một kiến trúc mới [17]. Tương tự transfomer, BERT có thể được học chuyển giao (transfer learning) và có thể được huấn luyện với các dữ liệu không cần gán nhãn. Hai kiểu huấn luyện BERT có thể được thực hiện đồng thời, đó là mặt nạ mô hình ngôn ngữ (Mask Language Model) và dự đoán câu kế tiếp (Next Sentence Prediction). BERT có thể được huấn luyện trước bằng một lượng lớn dữ liệu văn bản không gán nhãn để tạo ra một mô hình có tri thức tổng quát về các mối quan hệ giữa các từ và các câu. Sau đó, mô hình có thể được tinh chỉnh thêm (fine-tuned) bằng cách cho học chuyển giao trên các

[1] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

tập dữ liệu đặc thù, thường là các tập dữ liệu nhỏ hơn và có gán nhãn cho các tác vụ cụ thể. Quá huấn luyện trước và tinh chỉnh mô hình của BERT được mô tả trong hình số 2.



Hình 2. *Quá trình huấn luyện và tinh chỉnh BERT [17]*

Như đã nói ở trên, BERT sử dụng các transfomers, một kiến trúc có khả năng học các mối quan hệ giữa các từ sử dụng một cơ chế dựa trên sự tập trung (attention). Transfomer bao gồm một bộ mã hóa (encoder) có nhiệm vụ đọc các văn bản đầu vào. Nó cũng có một bộ giải mã (decoder) có nhiệm vụ dự đoán dựa trên tác vụ cần thực hiện. Khác với các mô hình theo cấu trúc một chiều, vốn đọc đầu vào theo thứ tự tuần tự, transfomer có khả năng đọc và xử lý tất cả các từ đầu vào cùng lúc và làm cho nó trở thành một mô hình có thể học ngữ cảnh của các tất cả các từ xung quanh theo cả hai chiều. Với kiến trúc đặc biệt đó và nhờ sử dụng một khối lượng khổng lồ dữ liệu huấn luyện, BERT đã cho kết quả tốt nhất trên 11 tác vụ phổ biến trong xử lý ngôn ngữ tự nhiên [17].

BERT có nhiều phiên bản được huấn luyện trước (pre-trained) cho các trường hợp sử dụng khác nhau. Hai phiên bản được sử dụng phổ biến nhất là BERT-base và BERT-large.

- BERT-base: Gồm 12 lớp mã hóa + 768 nút ẩn (hidden units) + 12 nút tập trung (attention heads). Tổng cộng 110 triệu tham số.
- BERT-large: Gồm 24 lớp mã hóa + 1024 nút ẩn (hidden units) + 12 nút tập trung (attention heads). Tổng cộng 340 triệu tham số.

2) PhoBERT

BERT là một mô hình đa ngôn ngữ, đã được huấn luyện và sử dụng cho nhiều ngôn ngữ khác nhau. Tuy nhiên, hầu hết các ngôn ngữ ngoài tiếng Anh đều được các nhóm nghiên cứu phát triển mô hình đặc thù cho ngôn ngữ đó dựa trên mô hình BERT ban đầu.

RoBERTa là một tiếp cận kế thừa kiến trúc và thuật toán của mô hình BERT nhưng mạnh và tối ưu hơn. Dự án này của Facebook hỗ trợ việc huấn luyện lại các mô hình BERT trên những bộ dữ liệu mới cho các ngôn ngữ khác ngoài một số ngôn ngữ phổ biến. Hiện đã có rất nhiều các mô hình huấn luyện trước cho những ngôn ngữ khác nhau được huấn luyện trên RoBERTa, trong đó có tiếng Việt với mô hình PhoBERT.

PhoBERT là một mô hình đơn ngôn ngữ cho tiếng Việt. PhoBERT dựa trên kiến trúc RoBERTa [18] và cho thấy hiệu suất tốt hơn hẳn so với các phương pháp dựa trên mô hình BERT đa ngôn ngữ khi làm việc với văn bản tiếng Việt.

3) Mô hình phân loại bài báo mạng dựa trên PhoBERT

Như đã nói ở trên, để sử dụng BERT hoặc PhoBERT cho bài toán phân loại, cần thực hiện tinh chỉnh mô hình đã huấn luyện trước trên các tập dữ liệu đặc thù đã được gán nhãn.

Trong nghiên cứu này, chúng tôi thực hiện tinh chỉnh mô hình PhoBERT trên tập dữ liệu phân loại tính chất bài báo đã được gán một trong ba nhãn tích cực, tiêu cực, và bình thường. Quá trình tinh chỉnh sử dụng thuật toán tối ưu AdamW, kỹ thuật tối ưu LayerNorm nhằm tìm kiếm kết quả tốt nhất cho bài toán phân loại. Mô hình thực hiện được mô tả như trong hình 3.

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. Dữ liệu và môi trường thực nghiệm

Trong nghiên cứu này, chúng tôi sử dụng tập dữ liệu được thu thập từ các trang báo điện tử tại Việt Nam từ năm 2020-2022. Mỗi bài báo được thu thập bao gồm thông tin về tiêu đề bài báo, tóm tắt bài báo và nội dung của bài báo. Nội dung của các bài báo đa dạng về các chủ đề y tế, chính trị, xã hội, môi trường v.v. Quá trình gán nhãn dữ liệu được hỗ trợ bởi các chuyên viên phòng Báo chí, Thành ủy Hà Nội. Việc hỗ trợ gán nhãn là việc cần thiết do quan điểm đánh giá các tin tức có tính chất tích cực hay tiêu cực cần dựa trên quan điểm cơ quan quản lý.

Tập dữ liệu huấn luyện bao gồm 6.364 bài báo được thu thập trong đó có 1.689 bài báo được gán nhãn tiêu cực, 2.608 bài báo được gán nhãn tích cực và 2.067 bài báo được gán nhãn bình thường, chúng tôi chia dữ liệu theo tỉ lệ 80-20 tương ứng với số dữ liệu huấn luyện và kiểm thử một cách ngẫu nhiên.

B. Độ đo đánh giá

Để đánh giá được kết quả mô hình phân loại, đầu tiên độ được sử dụng là Accuracy, độ đo này đánh giá chính xác kết quả dự đoán đúng hoặc sai của mô hình:

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

Trong đó TP (True Positive) là tổng số trường hợp dự báo khớp Positive, TN là tổng số trường hợp dự báo khớp với Negative.

Độ đo Presicion sẽ biểu diễn kết quả của các trường hợp dự báo là Positive thì có bao nhiêu trường hợp là đúng

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

Trong đó TP được trình bày ở công thức (1), FP là tổng số trường hợp dự báo các quan sát thuộc nhãn Negative thành Positive.

Độ đo Recall biểu diễn kết quả của các trường hợp Positive trên toàn bộ các mẫu thuộc nhóm Positive.

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

Trong đó TP được trình bày ở công thức (1), FN là tổng số trường hợp dự báo các quan sát thuộc nhãn Positive thành Negative.

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

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

C. Kết quả thực nghiệm

Mỗi bài báo khi thu thập gồm có ba phần là tiêu đề, tóm tắt, và nội dung, trong đó phần tiêu đề và tóm tắt mặc dù ngắn nhưng chứa các thông tin quan trọng về vụ việc. Chúng tôi tiến hành các thực nghiệm với các kịch bản khác nhau để đánh giá việc sử dụng phần nào của bài báo đem lại kết quả tốt nhất. Cụ thể các kịch bản phân loại bao gồm:

- Phân loại bằng tiêu đề + tóm tắt
- Phân loại bằng nội dung
- Phần loại bằng cả tiêu đề, tóm tắt, và nội dung

Ngoài ra, các thực nghiệm cũng được thực hiện với mô hình PhoBERT-base là PhoBERT-large để đánh giá hiệu năng cũng như hiệu suất của các mô hình.

Bảng 1 cho thấy kết quả của mô hình phân loại bài báo dựa trên trên các kịch bản. Các độ đo được sử dụng trong bảng bao gồm Accuracy, Precision, Recall, F1-score.

Theo kết quả trong bảng 1, sử dụng toàn bộ bài báo (bao gồm tiêu đều, tóm tắt, và nội dung) cho kết quả tốt nhất với độ chính xác trên mô hình PhoBERT-large là 93.1%. Nếu chỉ dùng tiêu đề + tóm tắt hoặc dùng nội dung thì kết quả khi dùng tiêu đề + tóm tắt là tốt hơn. Điều này có thể được lý giải do nội dung của bài báo có thể chứa nhiều từ hoặc câu gây nhiễu, ít liên quan tới tính chất tích cực hay tiêu cực của bài báo. Việc chỉ sử dụng tiêu đề + tóm tắt khi phân loại cho kết quả khá tốt cũng là một gợi ý cho các nghiên cứu tiếp theo. Lưu ý rằng tóm tắt được sử dụng trong thực nghiệm là phần đầu của mỗi bài báo chứ không phải đoạn tóm tắt được thực hiện bằng một kỹ thuật riêng rẽ.

Bảng 1. *Kết quả của mô hình dựa trên PhoBERT*

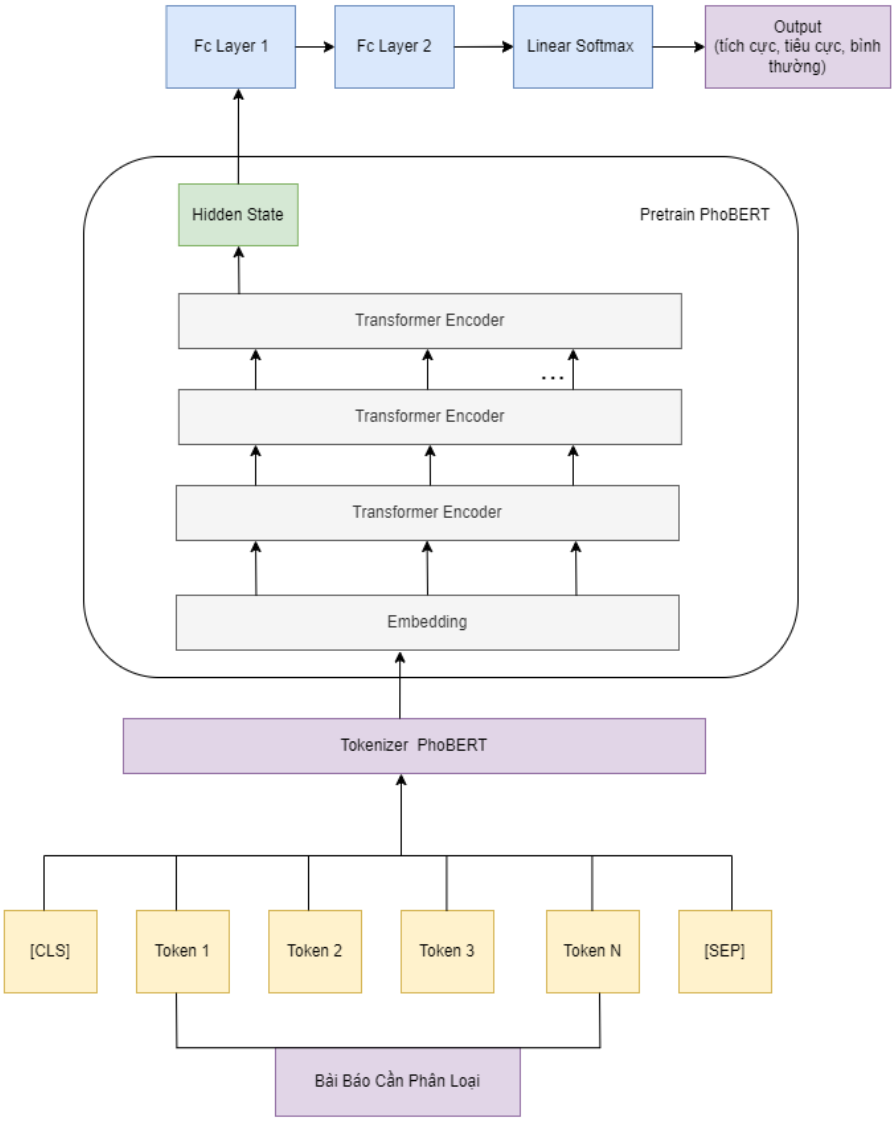
	Acc. (%)	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
Phân loại bằng tiêu đề + tóm tắt				
PhoBERT-base	91.6	91.9	92.1	92.0
PhoBERT-large	92.2	92.7	93.0	92.8
Phân loại bằng nội dung				
PhoBERT-base	90.5	91.2	90.7	90.9
PhoBERT-large	91.1	92.0	91.4	91.7
Phân loại bằng cả tiêu đề, tóm tắt, và nội dung				
PhoBERT-base	92.0	93.4	91.8	92.6
PhoBERT-large	**93.1**	**94.2**	**92.1**	**93.1**

Để đánh giá hiệu quả của mô hình dựa trên PhoBERT so với các cách tiếp cận khác như dựa trên mô hình BERT đa ngôn ngữ hay các mô hình sử dụng đặc trưng nhúng từ truyền thống, các thực nghiệm cũng được tiến hành sử dụng các phương pháp này tên cùng tập dữ liệu gồm cả tiêu đề, tóm tắt, và nội bài báo. Mô hình BERT đa ngôn ngữ cũng được sử dụng trong phân loại tương tự như PhoBERT, chỉ thay đổi ở phần mạng huấn luyện trước. Phương pháp dựa trên các đặc trưng nhúng từ sử dụng các véc tơ từ đã được huấn luyện trước trên tập dữ liệu tiếng Việt sử dụng kỹ thuật tạo véc tơ từ Word2Vec [20].

Các kết quả trong bảng 2 cho thấy các phương pháp dựa trên BERT có độ chính xác tốt hơn hẳn so với phương pháp truyền thống. Mô hình BERT đa ngôn ngữ mặc dù không được huấn luyện trước đặc thù cho tiếng Việt nhưng vẫn có kết quả tốt hơn mô hình Word2Vec đã được huấn luyện trước trên tiếng Việt (cao hơn khoảng 1%). Ngoài ra, kết quả trong bảng cũng khẳng định lại mô hình PhoBERT được huấn luyện trước đặc thù trên tiếng Việt đã cho kết quả tốt hơn mô hình BERT huấn luyện trước đa ngôn ngữ (cao hơn khoảng 2-3%).

Bảng 2. *So sánh kết quả theo các phương pháp*

Phương pháp	**Acc. (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
Pho-BERT large	**93.1**	**94.2**	**92.1**	**93.1**
BERT đa ngôn ngữ	90.2	92.7	89.3	90.6
Word2Vec	89.3	91.9	88.5	89.8



Hình 3. *Mô hình phân loại bài báo*

V. KẾT LUẬN

Nghiên cứu thực hiện phân loại bài báo mạng theo các tính chất tích cực, tiêu cực, hoặc bình thường sử dụng mô hình PhoBERT. Các thực nghiệm được thực hiện trên tập dữ liệu thu thập từ mạng Internet và quá trình gán nhãn được hỗ trợ bởi các chuyên viên quản lý thông tin mạng để đảm bảo chất lượng gán nhãn.

Các kết quả thực nghiệm cho thấy việc phân loại các bài báo mạng sử dụng mô hình mạng học sâu dựa trên BERT có nhiều tiềm năng áp dụng trong thực tế, làm giảm bớt nhân lực thủ công và tăng tốc độ thực hiện.

Hướng nghiên cứu có thể được phát triển trong tương lai với các bước tiền xử lý trước khi phân loại như xây dựng bản tóm tắt tự động nhằm giữ lại các thông tin chính và giảm bớt các thông tin nhiễu. Các bài cáo cũng có thể được phân loại theo chủ đề trước để làm cho việc phân loại theo tính chất được tập trung và cho kết quả tốt hơn.

TÀI LIỆU THAM KHẢO

[1] B.T. Kieu and S.B. Pham. Sentiment Analysis for Vietnamese. In Proceedings of Second International Conference on Knowledge and Systems Engineering (KSE), pp. 152–157, 2010

[2] N.T. Duyen, N.X. Bach, and T.M. Phuong, “An Empirical Study on Sentiment Analysis for Vietnamese”. The 2014 International Conference on Advanced Technologies for Communications (ATC’14)

[3] V.H. Quan, N.T. Huy, L. Bac, and N.L. Minh, “Multi-channel LSTM-CNN model for Vietnamese sentiment analysis”, 2017 9th International Conference on Knowledge and Systems Engineering (KSE).

[4] Q.T. Nguyen, T.L. Nguyen, N.H. Luong, and Q.H. Ngo, "Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews," 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh City, Vietnam, 2020, pp. 302-307, doi: 10.1109/NICS51282.2020.9335899.

[5] X. She and D. Zhang, "Text classification based on

hybrid CNN-LSTM hybrid model," in 2018 11th International Symposium on Computational Intelligence and Design (ISCID), 2018, vol. 2, pp. 185-189: IEEE.

[6] J. Cai, J. Li, W. Li, and J. Wang, "Deeplearning model used in text classification," in 2018 15th international computer conference on wavelet active media technology and information processing (ICCWAMTIP), 2018, pp. 123-126: IEEE.

[7] L. Lenc and P. Král, "Deep neural networks for Czech multi-label document classification," in International Conference on Intelligent Text Processing and Computational Linguistics, 2016, pp. 460-471: Springer.

[8] K. Kowsari, M. Heidarysafa, D. E. Brown, K. J. Meimandi, and L. E. Barnes, "Rmdl: Random multimodel deep learning for classification," in Proceedings of the 2nd international conference on information system and data mining, 2018, pp. 19-28.

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

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

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

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

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

[14] 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

[15] 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

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

[17] 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.

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

[19] N.Q.Dat and N.A.Tuan. PhoBERT: Pre-trained language models for Vietnamese. arXiv preprint arXiv:2003.00744. 2020.

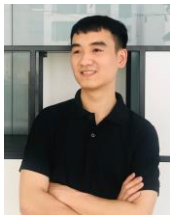
[20] V.Thanh, N.Q.Dat, N.Q.Dai, D.Mark, and J.Mark. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, NAACL 2018, pages 56-60.

VIETNAMESE NEWS ARTICLE CLASSIFICATION USING PhoBERT

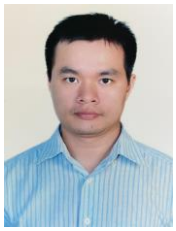
Abstract: Online information management is an important issue today when media content is increasingly digitized and popular. Due to the convenient nature of media and digital journalism, information about news stories is transmitted quickly and easily to readers. Along with this development, management agencies must also be able to quickly capture information for timely handling. In particular, quickly collecting and determining the nature of the news stories that is happening in the media is an important job for information management agencies. This paper proposes a system and method to automatically collect, evaluate and classify news articles through the content of online articles using the PhoBERT model. Experiments were conducted on a data set of more than 6,000 articles that were automatically collected and labeled by experts in the field of online information management. The highest classification results according to the accuracy and F1 measure to 93.1% according to the nature of the case with three labels of positive, negative, and normal, showing that the method is feasible and can be applied in practice.

Keywords: news articles classification, online information management, bert.

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 đang công tác tại Viện Tài nguyên Môi trường và Công nghệ thông tin (Inres.AI)



Nguyễn Đình Toàn Tốt nghiệp Kỹ sư Công nghệ Thông tin Học viện Công nghệ Bưu chính Viễn thông năm 2022. Hiện đang công tác tại Viện Tài nguyên Môi trường và Công nghệ thông tin (Inres.AI).



Lê Trường Thiên Tốt nghiệp Thạc sỹ chuyên ngành Hệ thống thông tin tại Đại học Công nghệ, Đại học Quốc gia Hà Nội năm 2003. Hiện đang công tác tại Viện Tài nguyên Môi trường và Công nghệ thông tin (Inres.AI).

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 đang công tác tại Khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông.