

So sánh hiệu quả của đặc trưng ngữ nghĩa PhoBERT và Naive Bayes trong phân loại chủ đề và đánh giá mức độ hài lòng từ bình luận của sinh viên

A Comparison of PhoBERT-Based and Naive Bayes Classification Methods for Topic Classification and Student Satisfaction Assessment in Student Feedback

Ngô Văn Hiếu^{a*}, Trịnh Quang Tín^b, Nguyễn Thị Thanh Phương^a, Nguyễn Thị Thuỳ Dung^a,
Trần Thị Thanh Lan^a, Nguyễn Phúc Minh Tú^a
Ngo Van Hieu^{a*}, Trinh Quang Tin^b, Nguyen Thi Thanh Phuong^a, Nguyen Thi Thuy Dung^a,
Tran Thi Thanh Lan^a, Nguyen Phuc Minh Tu^a

^aKhoa Công nghệ Thông tin, Trường Khoa học Máy tính và Trí tuệ Nhân tạo, Đại học Duy Tân, Đà Nẵng, Việt Nam

^aFaculty of Information Technology, School of Computer Science and Artificial Intelligence, Duy Tan University, Da Nang, 550000, Viet Nam

^bKhoa Khoa học Máy tính, Trường Khoa học Máy tính và Trí tuệ Nhân tạo, Đại học Duy Tân, Đà Nẵng, Việt Nam

^bFaculty of Computer Sciences, School of Computer Science and Artificial Intelligence, Duy Tan University, Da Nang, 550000, Viet Nam

(Ngày nhận bài: 29/12/2025, ngày phản biện xong: 23/01/2026, ngày chấp nhận đăng: 07/04/2026)

Tóm tắt

Bài báo này trình bày một nghiên cứu so sánh giữa mô hình học máy truyền thống NBC và phương pháp phân loại dựa trên PhoBERT trong bài toán phân tích phản hồi bằng tiếng Việt của sinh viên, bao gồm hai nhiệm vụ là phân loại chủ đề phản hồi và đánh giá mức độ cảm xúc. Tập dữ liệu được thu thập từ khảo sát phản hồi của sinh viên Đại học Duy Tân, trong đó các bình luận được gán nhãn theo ba chủ đề gồm giảng dạy, cơ sở vật chất và học phí, cùng năm mức độ cảm xúc từ rất không hài lòng đến rất hài lòng. Hai mô hình được huấn luyện và đánh giá trên cùng tập dữ liệu với quy trình và thước đo thống nhất nhằm đảm bảo tính công bằng. Kết quả thực nghiệm cho thấy cả hai mô hình đều đạt hiệu năng rất cao trong bài toán phân loại chủ đề, trong khi PhoBERT thể hiện ưu thế rõ rệt trong bài toán đánh giá mức độ cảm xúc với mức cải thiện F1-score từ khoảng 8% đến 10% so với NBC. Kết quả nghiên cứu cho thấy NBC phù hợp với các hệ thống yêu cầu tốc độ xử lý nhanh và chi phí thấp, trong khi PhoBERT là lựa chọn hiệu quả cho các ứng dụng phân tích cảm xúc chuyên sâu trong giáo dục đại học.

Từ khóa: phản hồi sinh viên, phân loại chủ đề, cảm xúc sinh viên, NBC, PhoBERT

Abstract

This paper presents a comparative study between the traditional machine learning model NBC and a PhoBERT-based classification approach for analyzing Vietnamese student feedback, focusing on two tasks: topic classification and sentiment level assessment. The dataset was collected from student feedback surveys at Duy Tan University, in which comments were labeled into three topics, including teaching, facilities, and tuition fees, and five sentiment levels ranging from very dissatisfied to very satisfied. Both models were trained and evaluated on the same dataset using identical procedures and evaluation metrics to ensure a fair comparison. Experimental results indicate that both models achieve very high performance in topic classification, while PhoBERT demonstrates a clear advantage in sentiment analysis, with F1-score improvements of approximately 8% to 10% compared to NBC. The results suggest that NBC is suitable for

*Tác giả liên hệ: Ngô Văn Hiếu
Email: ngovanhieu2@dtu.edu.vn

systems requiring fast processing and low computational cost, whereas PhoBERT is an effective choice for advanced sentiment analysis applications in higher education.

Keywords: student feedback, topic classification, student sentiment, NBC, PhoBERT

1. Giới thiệu

1.1. Đặt vấn đề

Trong giáo dục đại học, phản hồi của sinh viên là nguồn thông tin quan trọng nhằm đánh giá chất lượng đào tạo và mức độ hài lòng của người học. Thông qua các bình luận phản hồi, nhà trường có thể nhận diện những vấn đề liên quan đến hoạt động giảng dạy, cơ sở vật chất và chính sách học phí, từ đó đề xuất các giải pháp cải thiện phù hợp. Tuy nhiên, trong thực tế, phản hồi sinh viên chủ yếu được thể hiện dưới dạng văn bản tự do, ngắn gọn và đa dạng về cách diễn đạt, khiến việc phân tích thủ công gặp nhiều khó khăn, tốn thời gian và dễ mang tính chủ quan.

Để khắc phục những hạn chế này, các phương pháp xử lý ngôn ngữ tự nhiên và học máy đã được nghiên cứu và áp dụng nhằm tự động hóa quá trình phân tích phản hồi sinh viên. Các hướng tiếp cận này cho phép khai thác thông tin từ dữ liệu văn bản ở quy mô lớn, đồng thời hỗ trợ quá trình ra quyết định trong quản lý và nâng cao chất lượng đào tạo.

1.2. Giải quyết vấn đề

Trong số các phương pháp học máy truyền thống, Naive Bayes Classifier (NBC) [1] được sử dụng phổ biến trong các bài toán phân loại văn bản nhờ cấu trúc đơn giản, tốc độ huấn luyện nhanh và chi phí tính toán thấp. Để đạt hiệu quả phân loại, NBC được kết hợp với các kỹ thuật biểu diễn đặc trưng dựa trên thống kê tần suất từ, điển hình là Term Frequency–Inverse Document Frequency (TF–IDF) [2], nhằm chuyển đổi văn bản sang dạng đặc trưng số. Cách tiếp cận này cho kết quả tương đối tốt đối với các phản hồi ngắn và có cấu trúc đơn giản, tuy nhiên vẫn tồn tại hạn chế trong việc nắm bắt ngữ cảnh và các sắc thái cảm xúc phức tạp do giả định độc lập giữa các đặc trưng.

Song song với đó, sự phát triển của các mô hình học sâu dựa trên kiến trúc Transformer [3] đã mang lại những cải tiến đáng kể cho bài toán xử lý ngôn ngữ tự nhiên. Đối với tiếng Việt, PhoBERT [4] là một mô hình tiền huấn luyện tiêu biểu, cho phép học và biểu diễn ngữ nghĩa của văn bản thông qua cơ chế self-attention [5]. Trong đó, biểu diễn ngữ cảnh hai chiều được hiểu là khả năng mô hình mã hóa ý nghĩa của mỗi từ dựa đồng thời trên ngữ cảnh ở cả hai phía, bao gồm các từ đứng trước và các từ đứng sau trong câu. Nhờ vậy, mỗi từ được đặt trong mối quan hệ ngữ cảnh đầy đủ của toàn bộ câu, giúp mô hình hiểu chính xác hơn ý nghĩa và sắc thái cảm xúc của văn bản.

Khác với các phương pháp truyền thống như NBC, PhoBERT không yêu cầu thiết kế đặc trưng thủ công mà được tinh chỉnh trực tiếp trên dữ liệu huấn luyện, từ đó nâng cao khả năng phân loại đối với các phản hồi có cấu trúc ngôn ngữ phức tạp và các mức độ cảm xúc gần nhau. Tuy nhiên, ưu thế này đi kèm với yêu cầu cao hơn về tài nguyên tính toán và chi phí huấn luyện.

Mặc dù đã có nhiều nghiên cứu áp dụng riêng lẻ NBC hoặc các mô hình dựa trên PhoBERT cho phân tích phản hồi và cảm xúc, các nghiên cứu so sánh trực tiếp và toàn diện giữa hai hướng tiếp cận này trong bài toán phân loại đồng thời chủ đề phản hồi và mức độ hài lòng của bằng tiếng Việt của sinh viên vẫn còn hạn chế.

1.3. Đóng góp của bài báo

Xuất phát từ khoảng trống nghiên cứu nêu trên, bài báo này tập trung so sánh hiệu quả của NBC và PhoBERT trong phân loại chủ đề và đánh giá mức độ hài lòng từ bình luận của sinh viên tại Đại học Duy Tân. Cả hai mô hình được huấn luyện và đánh giá trên cùng một tập dữ liệu

khảo sát phản hồi sinh viên với quy trình tiên xử lý và các thước đo đánh giá thống nhất, nhằm đảm bảo tính công bằng trong so sánh.

Các đóng góp chính của bài báo bao gồm: (i) đề xuất một nghiên cứu so sánh có hệ thống giữa NBC, đại diện cho hướng tiếp cận học máy truyền thống dựa trên đặc trưng TF-IDF, và PhoBERT, đại diện cho hướng tiếp cận học sâu được tinh chỉnh trực tiếp trên dữ liệu để học biểu diễn ngữ cảnh; (ii) thực hiện đánh giá thực nghiệm công bằng trên cùng tập dữ liệu phản hồi sinh viên tại Đại học Duy Tân, trong đó các bình luận được gán nhãn theo ba chủ đề (giảng dạy, cơ sở vật chất và học phí), cũng như năm mức độ hài lòng (rất không hài lòng, không hài lòng, trung lập, hài lòng và rất hài lòng); (iii) phân tích và thảo luận định lượng và định tính về ưu điểm, hạn chế và các đánh đổi giữa hai mô hình, qua đó làm rõ mức độ cải thiện hiệu năng của PhoBERT so với NBC, đồng thời xem xét chi phí tính toán và khả năng triển khai trong các hệ thống phân tích phản hồi sinh viên trong môi trường giáo dục đại học.

2. Các nghiên cứu liên quan

2.1. Phân tích phản hồi sinh viên

Phân tích phản hồi sinh viên là một hướng nghiên cứu quan trọng trong lĩnh vực khai phá dữ liệu giáo dục, với mục tiêu hỗ trợ đánh giá chất lượng đào tạo và cải thiện trải nghiệm học tập của người học. Nhiều công trình đã khai thác dữ liệu khảo sát phản hồi sinh viên dưới dạng văn bản để thực hiện các nhiệm vụ như phân loại chủ đề phản hồi hoặc phân tích cảm xúc.

Các nghiên cứu trước đây cho thấy việc áp dụng các phương pháp xử lý ngôn ngữ tự nhiên giúp tự động hóa quá trình phân tích phản hồi sinh viên và giảm phụ thuộc vào đánh giá thủ công. Tuy nhiên, phần lớn các công trình này chỉ tập trung vào một nhiệm vụ cụ thể hoặc đánh giá hiệu quả của một mô hình riêng lẻ trên từng tập

dữ liệu, dẫn đến khó khăn trong việc so sánh và tổng quát hóa kết quả.

2.2. Các phương pháp học máy truyền thống trong phân tích văn bản

Các mô hình học máy truyền thống, trong đó NBC là đại diện tiêu biểu, đã được sử dụng rộng rãi trong các bài toán phân loại văn bản và phân tích cảm xúc. Khi kết hợp với các kỹ thuật biểu diễn đặc trưng như TF-IDF và n-gram [6], các mô hình này cho thấy hiệu quả ổn định đối với dữ liệu văn bản ngắn, bao gồm cả phản hồi sinh viên.

Mặc dù có ưu điểm về chi phí tính toán và khả năng triển khai, các nghiên cứu cũng chỉ ra rằng hiệu năng của các mô hình truyền thống suy giảm khi xử lý các văn bản có cấu trúc ngữ nghĩa phức tạp hoặc các bài toán phân loại đa mức độ cảm xúc. Những hạn chế này làm giảm khả năng ứng dụng của các mô hình học máy truyền thống trong các kịch bản đòi hỏi độ chính xác cao.

2.3. Các mô hình Transformer cho tiếng Việt

Sự phát triển của các mô hình học sâu dựa trên kiến trúc Transformer đã mang lại những cải tiến đáng kể cho xử lý ngôn ngữ tự nhiên. Đối với tiếng Việt, PhoBERT là một mô hình tiên huấn luyện được thiết kế chuyên biệt và đã được áp dụng thành công trong nhiều nhiệm vụ phân loại văn bản và phân tích cảm xúc.

Các nghiên cứu thực nghiệm cho thấy PhoBERT có khả năng khai thác ngữ cảnh và ngữ nghĩa tốt hơn so với các mô hình học máy truyền thống, đặc biệt trong các trường hợp văn bản có nội dung phức tạp hoặc ranh giới giữa các mức độ cảm xúc không rõ ràng. Tuy nhiên, hầu hết các công trình hiện có tập trung đánh giá PhoBERT một cách độc lập, trong khi các nghiên cứu so sánh trực tiếp với các mô hình học máy truyền thống trên cùng dữ liệu phản hồi sinh viên vẫn còn hạn chế.

2.4. Nhận xét và khoảng trống nghiên cứu

Tổng hợp các nghiên cứu liên quan cho thấy mỗi hướng tiếp cận đều có những ưu điểm và hạn chế riêng. Trong khi các mô hình học máy truyền thống phù hợp với các kịch bản hạn chế tài nguyên, các mô hình Transformer như PhoBERT cho thấy tiềm năng vượt trội về hiệu năng phân loại. Tuy nhiên, vẫn thiếu các nghiên cứu so sánh trực tiếp và công bằng giữa hai hướng tiếp cận này trong bài toán phân loại đồng thời chủ đề phản hồi và mức độ hài lòng từ các phản hồi bằng tiếng Việt của sinh viên. Khoảng trống này chính là động lực cho nghiên cứu được trình bày trong bài báo.

3. Phương pháp nghiên cứu

3.1. Thu thập dữ liệu

Dữ liệu sử dụng trong nghiên cứu này được thu thập thông qua một khảo sát trực tuyến được triển khai đến sinh viên Đại học Duy Tân vào cuối mỗi học kỳ. Mục tiêu của khảo sát là thu thập phản hồi của sinh viên liên quan đến trải nghiệm học tập trong môi trường giáo dục đại học, qua đó phục vụ cho bài toán phân tích phản hồi sinh viên bằng các phương pháp học máy và học sâu.

Nội dung khảo sát tập trung vào ba chủ đề chính gồm giảng dạy, cơ sở vật chất và học phí, phản ánh các khía cạnh cốt lõi trong quá trình đào tạo. Mỗi mẫu dữ liệu bao gồm hai thành phần thông tin: (i) phản hồi văn bản tự do của sinh viên, đóng vai trò là dữ liệu đầu vào; và (ii) các nhãn đầu ra tương ứng, bao gồm chủ đề phản hồi và mức độ hài lòng. Mức độ hài lòng được đánh giá theo thang đo năm mức gồm rất không hài lòng, không hài lòng, trung lập, hài lòng và rất hài lòng.

Ví dụ, một phản hồi văn bản đầu vào có nội dung “Giảng viên giảng bài dễ hiểu nhưng cơ sở vật chất của phòng học còn hạn chế” được gán nhãn đầu ra tương ứng gồm chủ đề giảng dạy và mức độ hài lòng hài lòng. Ví dụ này minh họa

cách dữ liệu phản hồi văn bản được liên kết với các nhãn phục vụ cho quá trình huấn luyện và đánh giá mô hình.

Ngoài ra, khảo sát còn thu thập một số thông tin cơ bản nhằm phục vụ cho việc xác thực và tổ chức dữ liệu ban đầu. Tuy nhiên, các thông tin định danh này không được sử dụng trong các giai đoạn phân tích tiếp theo. Trong quá trình tiền xử lý và huấn luyện mô hình, chỉ nội dung phản hồi văn bản và các nhãn tương ứng được giữ lại nhằm đảm bảo tính ẩn danh và bảo mật thông tin cá nhân. Tổng cộng hơn 3.000 phản hồi hợp lệ đã được thu thập và sử dụng làm dữ liệu đầu vào cho nghiên cứu.

3.2. Tiền xử lý và chuẩn hóa văn bản

Sau khi thu thập, dữ liệu phản hồi sinh viên được tiến hành tiền xử lý nhằm loại bỏ nhiễu và chuẩn hóa nội dung văn bản trước khi đưa vào huấn luyện mô hình. Trước hết, các phản hồi không hợp lệ, trùng lặp hoặc thiếu thông tin được loại bỏ khỏi tập dữ liệu. Văn bản phản hồi sau đó được chuẩn hóa về chữ thường, loại bỏ ký tự đặc biệt và các thành phần không mang nhiều ý nghĩa ngữ nghĩa.

Tiếp theo, dữ liệu được tách từ tiếng Việt [7] nhằm đảm bảo tính chính xác trong quá trình biểu diễn văn bản, trong đó các từ ghép nhiều âm tiết được xử lý như một đơn vị từ vựng thống nhất. Văn bản sau khi tách từ được sử dụng để xây dựng đặc trưng TF-IDF cho mô hình NBC và làm đầu vào cho bộ tokenizer [8] của PhoBERT trong giai đoạn fine-tuning [9].

Toàn bộ quy trình tiền xử lý được áp dụng thống nhất cho cả hai mô hình NBC và PhoBERT nhằm đảm bảo tính công bằng trong so sánh. Sau bước này, dữ liệu văn bản ở dạng chuẩn được sử dụng cho các giai đoạn gán nhãn, huấn luyện và đánh giá mô hình.

3.3. Gán nhãn và chia tập dữ liệu

Trong nghiên cứu này, cả hai mô hình NBC và PhoBERT đều được huấn luyện theo hướng

tiếp cận học có giám sát. Do đó, toàn bộ dữ liệu phản hồi sinh viên được gán nhãn trước khi đưa vào huấn luyện mô hình. Mỗi phản hồi được gán nhãn theo hai chiều thông tin, bao gồm chủ đề phản hồi và mức độ hài lòng, dựa trên nội dung khảo sát và phản hồi văn bản tương ứng.

Sau khi hoàn tất gán nhãn, tập dữ liệu được chia thành các tập con phục vụ cho huấn luyện và đánh giá mô hình. Việc chia dữ liệu được thực hiện thống nhất cho cả hai mô hình nhằm đảm bảo tính công bằng trong so sánh. Cụ thể, dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80% và 20%. Trong đó, tập huấn luyện được sử dụng để huấn luyện mô hình NBC và tinh chỉnh PhoBERT, trong khi tập kiểm tra được sử dụng để đánh giá hiệu năng phân loại. Đối với PhoBERT, một phần của tập huấn luyện được sử dụng làm tập validation [10] trong quá trình fine-tuning nhằm điều chỉnh siêu tham số và hạn chế hiện tượng overfitting.

3.4. Mô hình cơ sở: NBC

NBC được sử dụng như mô hình cơ sở nhằm làm mốc so sánh với PhoBERT trong nghiên cứu này. NBC là một mô hình phân lớp dựa trên xác suất, được xây dựng trên định lý Bayes và giả định các đặc trưng đầu vào là độc lập có điều kiện.

Với một phản hồi văn bản $d = \{w_1, w_2, \dots, w_n\}$ và một lớp c , xác suất để phản hồi d thuộc về lớp c được xác định theo định lý Bayes:

$$P(c | d) = \frac{P(d | c)P(c)}{P(d)} \quad (1)$$

Trong đó, xác suất hậu nghiệm $P(c | d)$ phản ánh mức độ phù hợp của phản hồi văn bản với lớp c . Dưới giả định độc lập có điều kiện giữa các đặc trưng, xác suất sinh ra văn bản d khi biết lớp c có thể được xấp xỉ bằng tích các xác suất xuất hiện của từng từ trong lớp đó:

$$P(d | c) \approx \prod_{i=1}^n P(w_i | c) \quad (2)$$

Do đó, quá trình phân lớp của NBC được thực hiện bằng cách so sánh các xác suất hậu nghiệm giữa các lớp và lựa chọn lớp có giá trị lớn nhất:

$$\hat{c} = \arg \max_{c \in C} P(c) \prod_{i=1}^n P(w_i | c) \quad (3)$$

Cách tiếp cận này cho phép NBC kết hợp thông tin xác suất tiên nghiệm của các lớp với mức độ xuất hiện của các đặc trưng trong văn bản, đồng thời giữ được hiệu quả tính toán nhờ giả định độc lập giữa các đặc trưng.

Trong nghiên cứu này, các phản hồi văn bản sinh viên sau khi tiền xử lý được tách từ tiếng Việt, trong đó các từ ghép nhiều âm tiết được xem như một đơn vị từ vựng nhằm bảo toàn ý nghĩa ngữ nghĩa.

Trước hết, tần suất xuất hiện của mỗi từ trong từng phản hồi được xác định và được xem là thành phần TF, phản ánh mức độ xuất hiện của từ đó trong văn bản. Tiếp theo, độ nghịch đảo tần suất văn bản, được gọi là IDF và được tính dựa trên số lượng phản hồi trong toàn bộ tập dữ liệu có chứa từ tương ứng, nhằm giảm trọng số của các từ xuất hiện phổ biến và làm nổi bật các từ mang tính phân biệt. Cuối cùng, trọng số TF-IDF của mỗi từ được xác định bằng cách kết hợp hai thành phần trên.

Trong nghiên cứu này, TF-IDF được xây dựng từ các n-gram với $n = 1, 2, 3$, bao gồm unigram, bigram và trigram. Do đó, mỗi phản hồi văn bản được biểu diễn dưới dạng một vector có số chiều bằng kích thước từ vựng, là tổng số unigram, bigram và trigram xuất hiện trong tập dữ liệu sau tiền xử lý. Để kiểm soát kích thước không gian đặc trưng khi sử dụng n-gram bậc cao, số lượng đặc trưng TF-IDF được giới hạn tối đa ở mức 15000 đặc trưng. Vì vậy, trong thực nghiệm, ma trận TF-IDF thu được có kích thước xấp xỉ $M \times 15000$, trong đó M là số lượng phản hồi trong tập huấn luyện.

Khi giá trị n trong n-gram tăng, số lượng đặc trưng tiềm năng cũng tăng theo, dẫn đến không gian đặc trưng có số chiều lớn và tính thưa cao. Tuy nhiên, trong thực nghiệm của bài báo, kích thước ma trận TF-IDF không tăng vô hạn theo n mà được khống chế thông qua giới hạn số đặc trưng và cơ chế loại bỏ các n-gram có tần suất xuất hiện thấp. Các vector TF-IDF sau đó được đưa vào mô hình NBC để huấn luyện và dự đoán nhãn, với cùng cách chia dữ liệu và quy trình đánh giá như các mô hình khác nhằm đảm bảo tính công bằng trong so sánh.

3.5. Mô hình đề xuất: PhoBERT

Trong nghiên cứu này, PhoBERT được sử dụng để xây dựng biểu diễn ngữ nghĩa của văn bản, làm đầu vào cho mô hình phân loại. Các phản hồi văn bản sau khi tiền xử lý được chuyển đổi thành chuỗi token thông qua bộ tokenizer tương ứng. Bộ tokenizer này thực hiện việc ánh xạ văn bản đầu vào sang dạng chuỗi token phù hợp với kiến trúc của PhoBERT, đồng thời tự động chèn các token đặc biệt theo quy định của mô hình.

Trong kiến trúc của PhoBERT, một token đặc biệt gọi là classification token (CLS) [11] được đặt ở đầu mỗi chuỗi đầu vào và được sử dụng để biểu diễn ngữ nghĩa tổng thể của toàn bộ phản hồi. Trong quá trình fine-tuning, PhoBERT đóng vai trò là bộ mã hóa ngữ cảnh, học biểu diễn ngữ nghĩa hai chiều của văn bản dựa trên cơ chế self-attention.

Để thực hiện bài toán phân loại, biểu diễn của token [CLS] ở đầu ra của PhoBERT được đưa qua một tầng phân loại fully connected (Dense) [12], sau đó áp dụng hàm Softmax [13] để suy ra xác suất của các lớp và dự đoán nhãn. Mô hình được huấn luyện theo hướng học có giám sát với hàm mất mát phù hợp cho bài toán phân loại đa lớp. Quá trình fine-tuning được thực hiện riêng cho hai nhiệm vụ, bao gồm phân loại chủ đề phản hồi và đánh giá mức độ hài lòng của sinh viên.

Khác với mô hình NBC, PhoBERT không yêu cầu thiết kế đặc trưng thủ công mà học trực tiếp biểu diễn ngữ nghĩa và ngữ cảnh từ dữ liệu huấn luyện. Nhờ đó, mô hình có khả năng xử lý hiệu quả các phản hồi có cấu trúc ngôn ngữ phức tạp cũng như phân biệt các mức độ cảm xúc gần nhau.

3.6. Thiết lập thực nghiệm và tiêu chí so sánh

Để đảm bảo tính công bằng trong so sánh, cả hai mô hình được huấn luyện và đánh giá trên cùng một tập dữ liệu, cùng nhãn và cùng các thước đo đánh giá. Quy trình tiền xử lý, cách chia dữ liệu và môi trường huấn luyện được giữ nhất quán cho tất cả các thí nghiệm. Các chỉ số đánh giá được sử dụng bao gồm Accuracy, Precision, Recall và F1-score [14].

Thiết lập này cho phép đánh giá khách quan sự khác biệt về hiệu năng phân loại cũng như chi phí tính toán giữa mô hình học máy truyền thống và mô hình học sâu trong bối cảnh phân tích phản hồi bằng tiếng Việt của sinh viên.

4. Thực nghiệm và kết quả

Phần này trình bày thiết lập thực nghiệm, các chỉ số đánh giá và kết quả so sánh hiệu năng giữa hai mô hình NBC và PhoBERT trong bài toán phân loại chủ đề phản hồi và đánh giá mức độ hài lòng từ bình luận của sinh viên Đại học Duy Tân. Toàn bộ các thí nghiệm được thiết kế nhằm đảm bảo tính công bằng và khả năng tái lập kết quả.

4.1. Thiết lập thực nghiệm

Các thí nghiệm trong nghiên cứu này được thực hiện trong hai môi trường tính toán khác nhau, phù hợp với đặc điểm và yêu cầu tài nguyên của từng mô hình.

Mô hình NBC được huấn luyện và đánh giá trên máy tính cá nhân sử dụng bộ xử lý Intel Core i7, bộ nhớ 16 GB RAM, chạy hệ điều hành Windows. Do đặc điểm của NBC có chi phí tính toán thấp, toàn bộ quá trình trích xuất đặc trưng, huấn luyện và đánh giá mô hình được thực hiện trên CPU mà không yêu cầu phần cứng chuyên dụng.

Đối với mô hình PhoBERT, các thí nghiệm fine-tuning được triển khai trên nền tảng Google Colaboratory Pro+, sử dụng bộ xử lý đồ họa NVIDIA Tesla T4 (16 GB VRAM). Môi trường thực nghiệm sử dụng Python 3.10, cùng các thư viện PyTorch 2.2 và Transformers. Việc sử dụng GPU giúp rút ngắn đáng kể thời gian huấn luyện và đảm bảo khả năng tinh chỉnh mô hình học sâu trên tập dữ liệu phản hồi sinh viên.

Dữ liệu phản hồi sinh viên sau khi được tiền xử lý và gán nhãn được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80% và 20%. Cách chia dữ liệu này được áp dụng thống nhất cho cả hai mô hình nhằm đảm bảo tính công bằng trong so sánh. Đối với mô hình PhoBERT, một phần của tập huấn luyện được sử dụng làm tập validation trong quá trình fine-tuning để điều chỉnh siêu tham số và hạn chế hiện tượng overfitting. Toàn bộ quá trình huấn luyện và đánh giá tuân thủ nguyên tắc không sử dụng dữ liệu kiểm tra trong bất kỳ giai đoạn huấn luyện nào, nhằm tránh rò rỉ dữ liệu và đảm bảo tính tổng quát của mô hình.

4.2. Bộ chỉ số đánh giá

Để đánh giá hiệu quả của các mô hình trong bài toán phân loại chủ đề phản hồi và mức độ hài lòng của sinh viên, nghiên cứu sử dụng bốn chỉ số đánh giá phổ biến gồm Accuracy, Precision, Recall và F1-score, được xác định như sau:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

Trong đó, các khái niệm True Positive (TP), False Positive (FP), True Negative (TN) và False Negative (FN) được xác định theo cách tiếp cận

một lớp so với phần còn lại cho từng nhiệm vụ phân loại như sau.

Đối với bài toán phân loại chủ đề phản hồi, bao gồm ba chủ đề là giảng dạy, cơ sở vật chất và học phí, mỗi chủ đề lần lượt được coi là lớp đang xét. Khi đó:

- TP: các nhãn được mô hình dự đoán đúng với nhãn thực tế là chủ đề đang xét.

- FP: các nhãn được mô hình dự đoán là chủ đề đang xét nhưng nhãn thực tế thuộc một trong các chủ đề còn lại.

- TN: các nhãn được mô hình dự đoán đúng với nhãn thực tế là không thuộc chủ đề đang xét, tức thuộc một trong các chủ đề còn lại.

- FN: các nhãn được mô hình dự đoán là một trong các chủ đề còn lại, nhưng nhãn thực tế lại là chủ đề đang xét.

Đối với bài toán đánh giá mức độ hài lòng, gồm năm mức độ là rất không hài lòng, không hài lòng, trung lập, hài lòng và rất hài lòng, cách xác định TP, FP, TN và FN cũng được thực hiện tương tự. Cụ thể, với mỗi mức độ hài lòng được xem là lớp đang xét. Khi đó:

- TP: các nhãn được mô hình dự đoán đúng với nhãn thực tế là mức độ hài lòng đang xét.

- FP: các nhãn được mô hình dự đoán là mức độ hài lòng đang xét nhưng nhãn thực tế thuộc một trong các mức độ còn lại.

- TN: các nhãn được mô hình dự đoán đúng với nhãn thực tế là không thuộc mức độ hài lòng đang xét, tức thuộc một trong các mức độ còn lại.

- FN: các nhãn được mô hình dự đoán là một trong các mức độ còn lại, nhưng nhãn thực tế lại là mức độ hài lòng đang xét.

Các chỉ số Precision, Recall và F1-score được tính theo từng lớp và tổng hợp cho bài toán phân loại đa lớp, còn Accuracy phản ánh mức độ chính xác tổng thể của mô hình trên toàn bộ dữ liệu.

4.3. Kết quả định lượng

Phần này trình bày kết quả so sánh định lượng giữa hai mô hình NBC và PhoBERT trong hai bài toán gồm phân loại chủ đề phản hồi và đánh giá

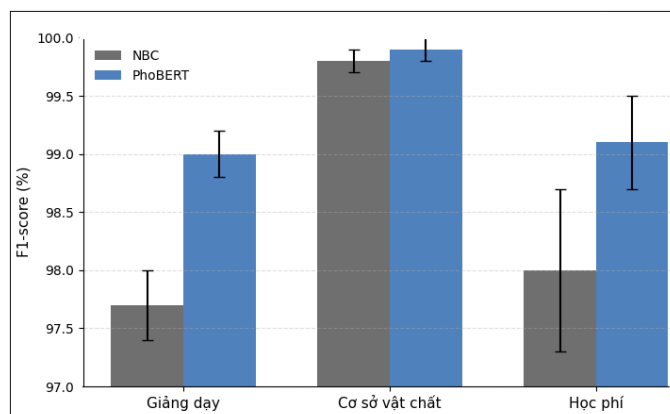
mức độ cảm xúc của sinh viên. Các kết quả được báo cáo dựa trên cùng tập dữ liệu kiểm tra và cùng bộ chỉ số đánh giá, nhằm đảm bảo tính công bằng trong so sánh hiệu năng giữa hai mô hình.

Bảng 1. Kết quả đánh giá phân loại theo chủ đề

Chủ đề phản hồi	Mô hình	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Giảng dạy	NBC	97.8 (±0.3)	97.7 (±0.3)	97.8 (±0.3)	97.7 (±0.3)
	PhoBERT	99.0 (±0.2)	99.1 (±0.2)	99.0 (±0.2)	99.0 (±0.2)
Cơ sở vật chất	NBC	99.8 (±0.1)	99.8 (±0.1)	99.8 (±0.1)	99.8 (±0.1)
	PhoBERT	99.9 (±0.1)	99.9 (±0.1)	99.9 (±0.1)	99.9 (±0.1)
Học phí	NBC	98.0 (±0.7)	98.0 (±0.7)	98.0 (±0.7)	98.0 (±0.7)
	PhoBERT	99.1 (±0.4)	99.2 (±0.4)	99.1 (±0.4)	99.1 (±0.4)

Bảng 1 trình bày kết quả phân loại theo chủ đề phản hồi, bao gồm ba chủ đề chính là giảng dạy, cơ sở vật chất và học phí. Nhìn chung, cả hai mô hình đều đạt hiệu năng rất cao trong bài toán này, với Accuracy và F1-score đều xấp xỉ

hoặc vượt mức 98%. Điều này cho thấy các chủ đề phản hồi có ranh giới ngữ nghĩa tương đối rõ ràng, giúp các mô hình học máy truyền thống như NBC vẫn có thể khai thác hiệu quả thông tin đặc trưng dựa trên TF-IDF và n-gram.



Hình 1. Biểu đồ so sánh F1-score của NBC và PhoBERT theo từng chủ đề phản hồi

Để trực quan hóa sự khác biệt về hiệu năng giữa hai mô hình, Hình 1 minh họa biểu đồ so sánh F1-score của NBC và PhoBERT theo từng chủ đề phản hồi. Biểu đồ cho thấy PhoBERT đạt hiệu năng nhỉnh hơn NBC trên cả ba chủ đề, tuy nhiên mức cải thiện là không lớn. Đối với chủ đề cơ sở vật chất, sự khác biệt giữa hai mô hình là

rất nhỏ do các phản hồi thường chứa các từ khóa mang tính đặc trưng cao. Trong khi đó, ở các chủ đề giảng dạy và học phí, PhoBERT thể hiện ưu thế rõ hơn nhờ khả năng khai thác ngữ cảnh tốt hơn, giúp mô hình xử lý hiệu quả các phản hồi có cách diễn đạt đa dạng và gián tiếp.

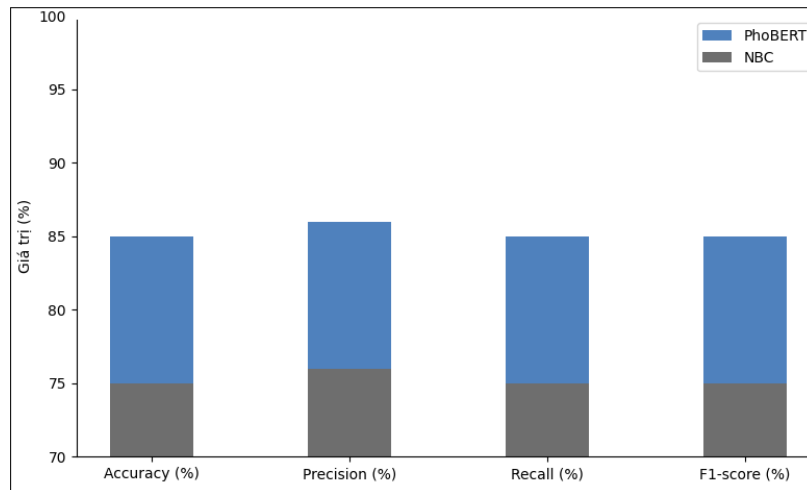
Bảng 2. Kết quả phân loại 5 mức cảm xúc theo từng chủ đề

Chủ đề phản hồi	Mô hình	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Giảng dạy	NBC	72.3 (±1.8)	72.2 (±1.9)	72.3 (±1.8)	72.1 (±1.9)
	PhoBERT	82.5 (±1.2)	82.8 (±1.1)	82.5 (±1.2)	82.6 (±1.1)
Cơ sở vật chất	NBC	80.3 (±1.1)	81.8 (±0.9)	80.3 (±1.1)	79.8 (±1.3)
	PhoBERT	88.1 (±0.8)	88.5 (±0.8)	88.1 (±0.8)	88.2 (±0.8)
Học phí	NBC	74.2 (±1.3)	74.0 (±1.3)	74.2 (±1.3)	74.0 (±1.3)
	PhoBERT	83.4 (±1.0)	83.7 (±1.0)	83.4 (±1.0)	83.5 (±1.0)

So với bài toán phân loại chủ đề, bài toán đánh giá mức độ cảm xúc cho thấy độ khó cao hơn đáng kể, thể hiện qua việc hiệu năng của mô hình NBC giảm rõ rệt, với F1-score dao động trong khoảng từ 72% đến 80% tùy theo chủ đề. Điều này phản ánh hạn chế của các mô hình học máy truyền thống trong việc nắm bắt sắc thái cảm xúc và ngữ cảnh ngôn ngữ phức tạp.

Ngược lại, PhoBERT đạt mức cải thiện đáng kể trên tất cả các chủ đề trong bài toán phân loại

cảm xúc. Mức tăng về F1-score so với NBC dao động từ khoảng 8% đến 10%, cho thấy khả năng vượt trội của PhoBERT trong việc học biểu diễn ngữ nghĩa theo ngữ cảnh. Đặc biệt, sự cải thiện rõ rệt ở các phản hồi liên quan đến giảng dạy và học phí cho thấy mô hình học sâu có lợi thế trong việc phân biệt các mức độ cảm xúc gần nhau, chẳng hạn như trung lập và hài lòng, vốn là những trường hợp khó đối với các phương pháp dựa trên đặc trưng thống kê truyền thống.



Hình 2. Biểu đồ so sánh F1-score của NBC và PhoBERT theo từng chủ đề phản hồi cảm xúc

Để làm rõ hơn mức cải thiện trong bài toán phân loại cảm xúc, Hình 2 trình bày biểu đồ so sánh F1-score của NBC và PhoBERT theo từng chủ đề phản hồi. Kết quả cho thấy khoảng cách hiệu năng giữa hai mô hình là rõ rệt và xuất hiện ở tất cả các chủ đề được xem xét, qua đó khẳng định ưu thế của PhoBERT trong các bài toán phân tích cảm xúc phản hồi sinh viên.

Tổng hợp các kết quả định lượng cho thấy PhoBERT chỉ mang lại mức cải thiện hạn chế trong bài toán phân loại chủ đề, nơi các ranh giới ngữ nghĩa tương đối rõ ràng, nhưng thể hiện ưu thế rõ rệt trong bài toán đánh giá mức độ cảm xúc, khẳng định hiệu quả của các mô hình Transformer trong việc khai thác ngữ cảnh và sắc thái ngôn ngữ của phản hồi bằng tiếng Việt của sinh viên.

5. Thảo luận

Bên cạnh các kết quả đánh giá định lượng đã trình bày ở phần trước, phần này tập trung thảo luận và phân tích định tính nhằm làm rõ sự khác biệt về hiệu năng giữa hai mô hình NBC và PhoBERT trong hai nhiệm vụ của bài toán phân tích phản hồi sinh viên, bao gồm phân loại chủ đề phản hồi và đánh giá mức độ hài lòng.

Đối với bài toán phân loại chủ đề phản hồi, kết quả thực nghiệm cho thấy cả NBC và PhoBERT đều đạt hiệu năng rất cao. Các chủ đề giảng dạy, cơ sở vật chất và học phí có ranh giới ngữ nghĩa tương đối rõ ràng và thường gắn với các từ hoặc cụm từ mang tính đặc trưng, giúp các mô hình dựa trên đặc trưng thống kê khai thác hiệu quả thông tin từ văn bản. Trong bối cảnh này, phần lớn các trường hợp dự đoán sai của

NBC xuất hiện ở những phản hồi có nội dung ngắn, cách diễn đạt gián tiếp hoặc không chứa nhiều từ khóa đặc trưng rõ ràng. Tuy nhiên, mức độ nhầm lẫn giữa các chủ đề là không lớn, do đặc thù ngữ nghĩa của từng nhóm phản hồi. PhoBERT xử lý tốt hơn các phản hồi có cấu trúc câu phức tạp hoặc đề cập đồng thời nhiều khía cạnh nhờ khả năng học biểu diễn ngữ cảnh hai chiều, song lợi thế này chỉ mang lại mức cải thiện nhẹ về hiệu năng, phù hợp với các kết quả định lượng đã quan sát.

Ngược lại, trong bài toán đánh giá mức độ hài lòng và cảm xúc của sinh viên, sự khác biệt giữa hai mô hình trở nên rõ rệt hơn. NBC thường gặp khó khăn khi phân biệt các mức độ cảm xúc gần nhau, đặc biệt là giữa trung lập và hài lòng hoặc giữa không hài lòng và rất không hài lòng. Nguyên nhân chính đến từ giả định độc lập có điều kiện giữa các đặc trưng và cách biểu diễn văn bản theo dạng bag-of-words kết hợp TF-IDF và n-gram, khiến mô hình chủ yếu dựa trên tần suất xuất hiện của từ hoặc cụm từ mà không nắm bắt được ngữ cảnh tổng thể của câu. Trong nhiều trường hợp, các mức độ cảm xúc lân cận được diễn đạt bằng những từ ngữ tương tự, trong khi sự khác biệt chủ yếu nằm ở sắc thái biểu đạt hoặc ngữ cảnh sử dụng, dẫn đến việc NBC khó xác định ranh giới giữa các lớp cảm xúc này.

Trong khi đó, PhoBERT thể hiện ưu thế rõ ràng trong bài toán đánh giá mức độ hài lòng nhờ khả năng khai thác ngữ cảnh và mối quan hệ giữa các từ trong câu. Mô hình có thể nắm bắt tốt hơn sắc thái cảm xúc tổng thể của phản hồi, ngay cả khi không xuất hiện các từ khóa cảm xúc rõ ràng. Điều này lý giải cho mức cải thiện đáng kể về các chỉ số Precision, Recall và F1-score của PhoBERT so với NBC trong bài toán phân loại cảm xúc, như đã thể hiện trong kết quả thực nghiệm.

Từ góc độ ứng dụng thực tế, các kết quả thảo luận cho thấy NBC vẫn là lựa chọn phù hợp cho các hệ thống phân tích phản hồi sinh viên khi mục tiêu chính là xác định nhanh chủ đề phản

hồi với chi phí tính toán thấp. Nhờ cấu trúc đơn giản và đặc trưng TF-IDF có chiều cố định, NBC có thể được triển khai hiệu quả trong các hệ thống khảo sát quy mô lớn, các công cụ tổng hợp ý kiến phục vụ báo cáo quản lý hoặc các dashboard theo dõi nhanh nội dung phản hồi theo chủ đề. Ngược lại, PhoBERT phù hợp hơn với các hệ thống phân tích chuyên sâu về mức độ hài lòng, nơi việc hiểu đúng ngữ cảnh và sắc thái ngôn ngữ đóng vai trò then chốt, chẳng hạn như các hệ thống giám sát chất lượng đào tạo, phát hiện sớm dấu hiệu bất mãn hoặc hỗ trợ đánh giá hiệu quả giảng dạy và dịch vụ sinh viên.

Xét tổng thể, nghiên cứu cho thấy không tồn tại một mô hình duy nhất tối ưu cho mọi mục tiêu ứng dụng. Việc lựa chọn mô hình cần dựa trên đặc thù của nhiệm vụ, yêu cầu về độ chính xác và khả năng đáp ứng tài nguyên tính toán. Một hướng tiếp cận thực tiễn là sử dụng NBC cho các bài toán phân loại chủ đề nhằm đảm bảo tốc độ xử lý và chi phí thấp, đồng thời áp dụng PhoBERT cho các bài toán đánh giá mức độ hài lòng yêu cầu khả năng hiểu ngữ cảnh và sắc thái ngôn ngữ cao hơn. Cách tiếp cận này cho phép tận dụng ưu điểm của cả hai mô hình và phù hợp với điều kiện triển khai thực tế trong môi trường giáo dục đại học.

6. Kết luận và hướng phát triển

Bài báo đã trình bày nghiên cứu so sánh giữa mô hình học máy truyền thống NBC và phương pháp phân loại dựa trên PhoBERT trong bài toán phân tích phản hồi bằng tiếng Việt của sinh viên, bao gồm hai nhiệm vụ là phân loại chủ đề phản hồi và đánh giá mức độ hài lòng. Các thí nghiệm được thiết kế công bằng trên cùng một tập dữ liệu phản hồi sinh viên nhằm đánh giá khách quan hiệu năng của hai mô hình. Kết quả thực nghiệm cho thấy cả NBC và PhoBERT đều đạt hiệu năng cao trong bài toán phân loại chủ đề phản hồi, với sự khác biệt không đáng kể giữa hai phương pháp. Điều này cho thấy các mô hình học máy truyền thống vẫn có thể đáp ứng tốt các

nhệm vụ có ranh giới ngữ nghĩa rõ ràng với chi phí tính toán thấp. Ngược lại, trong bài toán đánh giá mức độ cảm xúc, PhoBERT thể hiện ưu thế rõ rệt nhờ khả năng khai thác ngữ cảnh và sắc thái ngôn ngữ của văn bản, mang lại sự cải thiện đáng kể về các chỉ số đánh giá so với NBC. Từ góc độ ứng dụng, kết quả nghiên cứu nhấn mạnh rằng việc lựa chọn mô hình phân tích phản hồi sinh viên cần gắn liền với mục tiêu cụ thể của hệ thống. NBC phù hợp với các hệ thống yêu cầu tốc độ xử lý nhanh và chi phí triển khai thấp, trong khi PhoBERT là lựa chọn hiệu quả cho các hệ thống phân tích cảm xúc chuyên sâu đòi hỏi độ chính xác cao.

Trong các nghiên cứu tiếp theo, tập dữ liệu phản hồi sinh viên có thể được mở rộng về quy mô và phạm vi, bao gồm nhiều khóa học và bối cảnh đào tạo khác nhau, nhằm nâng cao khả năng tổng quát hóa của mô hình. Bên cạnh đó, việc thử nghiệm và so sánh với các mô hình Transformer khác cho tiếng Việt hoặc các mô hình đa ngôn ngữ cũng là một hướng nghiên cứu tiềm năng. Ngoài ra, nghiên cứu trong tương lai có thể tập trung vào việc tích hợp các mô hình phân tích phản hồi vào các hệ thống quản lý đào tạo thực tế, kết hợp với dữ liệu học tập và hành vi sinh viên để hỗ trợ phát hiện sớm các vấn đề trong quá trình đào tạo. Cuối cùng, việc nghiên cứu các chiến lược tối ưu hóa nhằm giảm chi phí tính toán của các mô hình học sâu sẽ góp phần nâng cao khả năng triển khai các hệ thống phân tích phản hồi sinh viên trong môi trường giáo dục đại học.

Tài liệu tham khảo

- [1] Zhang, L. (2025). "Features extraction based on Naive Bayes algorithm and TF-IDF for news classification". *PLoS ONE* (20), e0327347. DOI: 10.1371/journal.pone.0327347.
- [2] Ionendri, N.A., Candra, F., & Rizal, A. (2025). "News Classification using Natural Language Processing with TF-IDF and Multinomial Naïve Bayes". *Journal of Applied Computer Science and Technology* (6), 37–45. DOI: 10.52158/jacost.v6i1.1099.
- [3] Rink, L., Meijdam, J., & Graus, D. (2024). Aspect-Based Sentiment Analysis for Open-Ended HR Survey Responses. *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, St. Julian's, Malta, 2024 (tr. 16–26). Stroudsburg: Association for Computational Linguistics.
- [4] Ngo, T.M., Ngo, B.H., & Stuchilin, V.V. (2024). "Fine-tuned PhoBERT for sentiment analysis of Vietnamese phone reviews". *CTU Journal of Innovation and Sustainable Development* (16), 52–57. DOI: 10.22144/ctujoisd.2024.322.
- [5] Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). "Self-attention attribution: Interpreting information interactions inside transformer". *Proceedings of the AAAI Conference on Artificial Intelligence* (35), 12963–12971.
- [6] Van, H.N., Phan, L., Nhat, V.T., & Viet, N.H.L. (2025). "A system for analyzing topics and evaluating satisfaction levels from Vietnamese student feedback using Naive Bayes classifier". *International Journal on Integrating Technology in Education* (14). DOI: 10.5121/ijite.2025.14304. Ghi chú: số trang mình thấy nguồn tìm kiếm đang hiển thị chưa hoàn toàn thống nhất, nên nên mở PDF gốc để chốt chính xác trước khi nộp.
- [7] Vu, T., Nguyen, D.Q., Nguyen, D.Q., Dras, M., & Johnson, M. (2018). VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, New Orleans, Louisiana, 2018* (tr. 56–60). Stroudsburg: Association for Computational Linguistics. DOI: 10.18653/v1/N18-5012.
- [8] Nguyen, D.Q., & Nguyen, A.T. (2020). PhoBERT: Pre-trained language models for Vietnamese. *Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020* (tr. 1037–1042). Stroudsburg: Association for Computational Linguistics. DOI: 10.18653/v1/2020.findings-emnlp.92.
- [9] Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018* (tr. 328–339). Stroudsburg: Association for Computational Linguistics. DOI: 10.18653/v1/P18-1031.
- [10] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995* (tr. 1137–1143).
- [11] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Minneapolis, Minnesota*,

2019 (tr. 4171–4186). Stroudsburg: Association for Computational Linguistics.

- [12] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.
- [13] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer. Mục này mình chưa mở được nguồn xuất bản chính

thức có đầy đủ chi tiết thành phố trong kết quả tìm kiếm hiện tại, nhưng cách ghi này là chuẩn bibliographic thông dụng của sách này.

- [14] Sokolova, M., & Lapalme, G. (2009). “A systematic analysis of performance measures for classification tasks”. *Information Processing & Management* (45), 427–437. DOI: 10.1016/j.ipm.2009.03.002.