Journal of Finance – Marketing Research; Vol. 16, Issue 1; 2025
p-ISSN: 1859-3690; e-ISSN: 3030-427X
DOI: https://doi.org/10.52932/jfm.v16i1

Journal of Finance - Marketing Research
http://jfm.ufm.edu.vn

SENTIMENT ANALYSIS ON STUDENT FEEDBACK USING BERT COMBINED WITH MULTI-CHANNEL CNN-GRU ARCHITECTURE

Tran Son Nam[1]*, Thai Kim Phung[1], Pham The Vinh[1]
[1]University of Economics Ho Chi Minh City, Vietnam

ARTICLE INFO

DOI:
10.52932/jfm.v16i1.611

Received:
September 09, 2024
Accepted:
November 19, 2024
Published:
February 25, 2025

Keywords:
Deep learning; Education management; Natural language processing; Sentiment analysis

JEL codes:
C61; C63; C67

ABSTRACT

Student feedback is a valuable data source for enhancing teaching quality and improving learner satisfaction. Numerous studies have conducted sentiment analysis on this data, yielding notable results. However, research in the Vietnamese language still faces significant limitations, including a limited number of published studies, challenges related to the target of sentiment, and data issues such as imbalance that pose difficulties for application. This study proposes a model that combines BERT with a multi-channel architecture consisting of CNN and GRU. By leveraging the strengths of each network, the performance of sentiment analysis on Vietnamese student feedback is expected to improve. The model focuses on classification tasks (topic and sentiment polarity) and supporting specific satisfaction measurements. Additionally, the model's ability to handle data imbalance is emphasized to utilize available datasets, saving time and finance effectively. Experiments on the UIT-VSFC dataset show performance improvements in Macro F1-Score compared to recent studies, with an increase of 0,01 in the topic classification task and 0,0051 in the sentiment polarity task. The study's result will be a useful solution for educational institutions, which can be applied to improve teaching, reputation management, and learner support and be a motivation for expanding future research.

*Corresponding author:
Email: namts@ueh.edu.vn

Tạp chí Nghiên cứu Tài chính - Marketing
http://jfm.ufm.edu.vn

PHÂN TÍCH CẢM XÚC TRÊN PHẢN HỒI HỌC VIÊN BẰNG MÔ HÌNH BERT KẾT HỢP KIẾN TRÚC ĐA KÊNH CNN-GRU

Trần Sơn Nam[1]*, Thái Kim Phụng[1], Phạm Thế Vinh[1]
[1]Đại học Kinh tế Thành phố Hồ Chí Minh

THÔNG TIN

DOI:
10.52932/jfm.v16i1.611

Ngày nhận:
09/09/2024
Ngày nhận lại:
19/11/2024
Ngày đăng:
25/02/2025

Từ khóa:
Học sâu;
Phân tích cảm xúc;
Quản lý giáo dục; Xử lý ngôn ngữ tự nhiên

Mã JEL:
C61; C63; C67

TÓM TẮT

Phản hồi học viên là một trong những nguồn dữ liệu quý giá để nâng cao chất lượng giảng dạy và cải thiện sự hài lòng người học. Nhiều nghiên cứu về phân tích cảm xúc trên nguồn dữ liệu này đã được thực hiện và mang lại các kết quả đáng ghi nhận. Tuy nhiên, các nghiên cứu trên ngôn ngữ tiếng Việt vẫn còn nhiều hạn chế, liên quan đến số lượng nghiên cứu công bố, mục tiêu cảm xúc hay vấn đề về dữ liệu như mất cân bằng gây khó khăn khi ứng dụng. Nghiên cứu này đề xuất một mô hình kết hợp BERT và kiến trúc đa kênh gồm CNN và GRU. Bằng việc tận dụng ưu điểm từng mạng, hiệu suất bài toán phân tích cảm xúc trên phản hồi học viên tại Việt Nam được kỳ vọng nâng cao. Trong đó, mô hình tập trung cả hai nhiệm vụ phân loại (chủ đề và cực cảm xúc), hỗ trợ đo lường sự hài lòng cụ thể. Đồng thời, khả năng chống mất cân bằng của mô hình được chú trọng nhằm khai thác hiệu quả các bộ dữ liệu sẵn có, giúp tiết kiệm thời gian và tài chính. Thực nghiệm trên bộ dữ liệu UIT-VSFC cho thấy sự cải thiện hiệu suất tại chỉ số F1-Score (Macro) so với nghiên cứu gần đây, tăng 0,01 và 0,0051 lần lượt tại nhiệm vụ chủ đề và cực cảm xúc. Kết quả của nghiên cứu sẽ là một giải pháp hữu ích cho các cơ sở giáo dục, có thể ứng dụng để cải thiện giảng dạy, quản lý danh tiếng, hỗ trợ người học và là động lực để mở rộng nghiên cứu trong tương lai.

1. Giới thiệu

Đối với các cơ sở giáo dục, việc nâng cao chất lượng giảng dạy luôn là vấn đề được đặc biệt quan tâm, với nhiều thành phần tham gia. Trong đó, vai trò của học viên được công nhận với yếu tố đánh giá là sự hài lòng người học (Razinkina và cộng sự, 2018). Yếu tố này thường được thể hiện thông qua những phản hồi, mang ý kiến chủ quan về các khía cạnh của cơ sở giáo dục mà học viên được tiếp xúc trong quá trình học tập, như giáo viên giảng dạy, nội dung chương trình học, cơ sở vật chất,... Vì vậy, việc khai thác phản hồi từ học viên sẽ giúp các cơ sở giáo dục có cơ hội để hiểu những điểm mạnh và điểm yếu đang tồn tại. Từ đó, những hành động duy trì và cải tiến phù hợp sẽ được thực hiện nhằm cải thiện mức độ hài lòng người học, góp phần nâng cao chất lượng giảng dạy.

**Tác giả liên hệ:*
Email: namts@ueh.edu.vn

Phân tích cảm xúc hay Sentiment analysis (SA) là một phương pháp khai thác phản hồi dạng văn bản và là một lĩnh vực quan trọng của Xử lý ngôn ngữ tự nhiên hay Natural language processing (NLP). Với nhiều ứng dụng, SA được thực hiện nhằm xác định cực cảm xúc (tích cực, tiêu cực hoặc trung tính) đối với thực thể theo quan điểm con người (Liu, 2022).

Hiện nay, các nhà nghiên cứu đã và đang áp dụng SA trên phản hồi học viên theo nhiều hướng tiếp cận khác nhau và đạt được những kết quả đáng ghi nhận (Shaik và cộng sự, 2023). Tuy nhiên, các nghiên cứu SA trên phản hồi học viên tiếng Việt vẫn còn tồn tại một số hạn chế. Thứ nhất, số lượng nghiên cứu tương đối ít so với các ngôn ngữ khác như tiếng Anh hoặc tiếng Trung (Kastrati và cộng sự, 2021), tạo ra khoảng trống về tối ưu hóa hiệu suất. Thứ hai, nhiều nghiên cứu chỉ tập trung phân loại cực cảm xúc và bỏ qua mục tiêu cảm xúc, khiến việc xác định sự hài lòng không cụ thể. Thứ ba, số lượng bộ dữ liệu về phản hồi tiếng Việt của học viên được công bố vẫn còn ít và hoạt động xây dựng một bộ dữ liệu mới phục vụ đào tạo mô hình tiêu tốn nhiều thời gian và tài chính. Thứ tư, UIT-VSFC (Nguyen Van Kiet và cộng sự, 2018) là bộ dữ liệu duy nhất hiện nay về phản hồi tiếng Việt của học viên, cho phép truy cập miễn phí nhưng bị mất cân bằng (Duong Vu Xuan Quynh và cộng sự, 2021), gây khó khăn trong đào tạo mô hình. Từ những hạn chế trên, yêu cầu về một mô hình SA tiếng Việt mới được đặt ra, không chỉ cải thiện hiệu suất mà còn có khả năng phân loại cảm xúc với mục tiêu cụ thể. Bên cạnh, mô hình cũng cần khai thác tối ưu bộ dữ liệu sẵn có, cụ thể là UIT-VSFC nhằm tiết kiệm chi phí. Việc đáp ứng các yêu cầu sẽ là mục đích mà nghiên cứu này hướng đến.

Gần đây, Học sâu hay Deep learning đã trở thành hướng tiếp cận phổ biến trong các nghiên cứu SA, liên quan các mạng, như: Long Short-term Memory (LSTM), Gated Recurrent Units (GRU), Convolutional Neural Network (CNN) và các biến thể khác. Với đặc điểm khác nhau, mỗi mạng có những ưu điểm riêng biệt. Để đạt hiệu quả cao, Vo Hoang Quan và cộng sự (2017) đề xuất một kiến trúc đa kênh nhằm trích xuất thông tin từ dữ liệu bằng cách đặt các mạng song song, và đạt hiệu suất vượt trội so với các mạng riêng lẻ. Ngoài ra, sự xuất hiện của Bidirectional Encoder Representations from Transformers (BERT) được đề xuất bởi Devlin và cộng sự (2019) cũng đã thúc đẩy nghiên cứu SA trên toàn thế giới nhờ hiệu quả biểu diễn từ mà mô hình này mang lại (Alaparthi & Mishra, 2021).

Để giải quyết các hạn chế, nghiên cứu này đề xuất một mô hình kết hợp mới theo hướng học sâu, bao gồm hai thành phần: BERT và kiến trúc đa kênh. Trong đó, thành phần BERT sử dụng mô hình PhoBERT (Nguyen Quoc Dat & Nguyen Tuan Anh, 2020) như một lớp nhúng, chuyển đổi đầu vào thành vectơ. Sau đó, các vectơ sẽ được đưa sang kiến trúc đa kênh, để nắm bắt thông tin cục bộ của các từ lân cận với CNN và nắm bắt thông tin tổng thể của toàn bộ chuỗi với GRU (Cho và cộng sự, 2014). Các thông tin sẽ kết hợp lại và đưa vào lớp phân loại chủ đề và cực cảm xúc. Hoạt động đào tạo và thực nghiệm mô hình sẽ tiến hành trên bộ dữ liệu sẵn có UIT-VSFC để giảm chi phí xây dựng bộ dữ liệu mới. Nhằm cải thiện hiệu suất và khắc phục vấn đề của UIT-VSFC, nâng cao chỉ số F1-Score (Macro) sẽ là mục tiêu của nghiên cứu này.

Kết quả đạt được sẽ đóng góp về nghiên cứu khoa học nói chung và vào lĩnh vực SA trên phản hồi học viên tiếng Việt nói riêng. Trong đó, các hạn chế hiện nay thuộc lĩnh vực nghiên cứu hướng đến sẽ được giải quyết hiệu quả với một mô hình đáp ứng mục đích cũng như mục tiêu đề ra. Điều này giúp tháo gỡ các nút thắt, lấp đầy khoảng trống nghiên cứu và tạo điều kiện thúc đẩy các nghiên cứu thuộc phạm vi của lĩnh vực giáo dục trong tương lai. Đồng thời, kết quả đạt được của nghiên cứu là một mô hình được xem như một giải pháp tối ưu về chi phí xây dựng và mang tính ứng dụng đối với các cơ sở giáo dục, giúp gia tăng giá trị cho đơn vị thông qua việc hỗ trợ các bài toán trong quản lý giáo dục như cải tiến chất lượng giảng dạy, quản lý danh tiếng, chấm điểm chuẩn các cơ sở hay hỗ trợ cho người học.

2. Các nghiên cứu liên quan

Theo Liu (2022), SA còn được gọi là khai thác ý kiến hay opinion mining, là lĩnh vực nghiên cứu về ý kiến, được thể hiện với năm thành phần gồm thực thể mục tiêu (sản phẩm, dịch vụ, sự kiện,…), khía cạnh hay thuộc tính của thực thể (chất lượng, học phí, công tác hỗ trợ,…), cảm xúc đối với khía cạnh được đề cập (thường là tích cực, tiêu cực hoặc trung tính), người nắm giữ hay thể hiện ý kiến và thời gian thể hiện ý kiến.

$$\textit{Ý kiến}\ (e, a, s, h, t) \quad (1)$$

Điều này dẫn đến việc SA bao gồm các nhiệm vụ cần được thực hiện, liên quan đến việc trích xuất và phân loại thực thể, khía cạnh và người nắm giữ ý kiến, trích xuất và chuẩn hóa thời gian, hồi quy hoặc phổ biến là phân loại cảm xúc. Trong đó, các nhiệm vụ hỗ trợ cho thành phần thực thể, khía cạnh và cảm xúc là quan trọng nhất. Một số trường hợp thực tế, các khía cạnh có thể bị bỏ qua mà chỉ tập trung vào thực thể, dẫn đến việc cắt giảm nhiệm vụ trích xuất và phân loại khía cạnh trong SA. Ngoài ra, người nắm giữ ý kiến và thời gian thể hiện ý kiến cũng có thể bị bỏ qua bởi các thành phần này có thể dễ dàng thu thập theo cấu trúc yêu cầu thông qua hệ thống và các cá nhân cần được ẩn danh trong quá trình thể hiện ý kiến. Với đặc điểm của các nhiệm vụ, SA đã tạo ra nhu cầu ứng dụng trong nhiều lĩnh vực khác nhau như khách sạn, hàng không, chăm sóc sức khỏe, chứng khoán,… nhằm phân tích, nâng cao mức độ hài lòng của người dùng dựa trên đa dạng các loại dữ liệu khác nhau, tập trung chủ yếu ở dữ liệu văn bản (Wankhade và cộng sự, 2022). Để thực hiện các nhiệm vụ, nhiều hướng tiếp cận đã được sử dụng trong các nghiên cứu.

Học sâu là một nhánh đặc biệt và mới nổi của học máy, áp dụng hướng tiếp cận học tập dựa trên nhiều lớp để đạt được những hiểu biết tốt nhất về dữ liệu. Các mạng nơ-ron là những hiện diện của học sâu, được sử dụng trong việc học tập dữ liệu (Chollet, 2021). Dựa trên đặc điểm thiết kế, học sâu bao gồm nhiều loại mạng nơ-ron khác nhau: CNN, RNN, LSTM, GRU, Transformer,… Theo đó, CNN là một loại mạng nơ-ron truyền thẳng, chuyên dụng để xử lý dữ liệu có cấu trúc dạng lưới, thường ứng dụng trong thị giác máy tính. Về kiến trúc, CNN điển hình sử dụng các lớp tích chập (convolution) chứa một hoặc nhiều bộ lọc (filter/kernel) có khả năng trượt trên dữ liệu để tính toán biểu diễn đặc trưng và đưa qua lớp gộp (pooling) để điều chỉnh độ phức tạp của biểu diễn trước khi sử dụng cho các mục đích khác (Goodfellow và cộng sự, 2016). Đối với LSTM, đây là mạng được giới thiệu bởi Hochreiter và Schmidhuber (1997) nhằm thay thế cho Recurrent Neural Network (RNN) vốn gặp nhiều hạn chế, sử dụng rộng rãi cho nhiều nhiệm vụ mô hình hóa chuỗi. So với RNN, LSTM có ba cổng tính toán hỗ trợ với cổng đầu vào (input gate) kiểm soát việc thông tin mới nào sẽ được lưu, cổng quên (forget gate) kiểm soát thông tin nào sẽ được loại bỏ và cổng đầu ra (output gate) xác định thông tin nào sẽ được sử dụng. Phát triển những năm gần đây, GRU là một biến thể của RNN do Cho và cộng sự (2014) đề xuất và dựa trên LSTM nhưng có cấu trúc tinh gọn và ưu điểm tốc độ, phù hợp với khối lượng dữ liệu lớn. Sự tinh gọn của GRU được thể hiện qua việc cắt giảm số lượng cổng với chỉ cổng đặt lại (reset gate) quyết định mức độ thông tin cũ cần loại bỏ và cổng cập nhật (update gate) kiểm soát mức độ sử dụng của thông tin cũ và mới. Nổi bật nhất, Transformer của Vaswani (2017) là một mạng nơ-ron tiên tiến, có khả năng tính toán song song hỗ trợ tối ưu tốc độ và sử dụng cơ chế self-attention để quan sát sự liên quan với các từ khác trong quá trình biểu diễn một từ. Với hai thành phần chính, bộ mã hóa (encoder) nhận dữ liệu đầu vào và tạo ra một biểu diễn vectơ có kích thước cố định, sau đó được đưa vào bộ giải mã (decoder) để tạo dữ liệu đầu ra. Một trong những mô hình dựa trên Transformer nổi tiếng nhất là BERT.

Trên thế giới, lĩnh vực Trí tuệ nhân tạo và NLP đã có sự phát triển vượt bật những năm gần đây. Điều này khuyến khích các nghiên cứu SA về dữ liệu giáo dục hay phản hồi sinh viên, với ngôn ngữ phổ biến là tiếng Anh và tiếng Trung (Kastrati và cộng sự, 2021). Nhiều hướng

tiếp cận đã triển khai, như: dựa trên từ vựng, dựa trên ngữ liệu, học máy và học sâu (Shaik và cộng sự, 2023). Trong đó, học sâu là hướng tiếp cận nhận được nhiều sự quan tâm với số lượng nghiên cứu ngày càng gia tăng. Dưới đây sẽ là một số nghiên cứu SA ứng dụng vào lĩnh vực giáo dục theo hướng tiếp cận học sâu trên thế giới những năm gần đây.

Đối với nhóm nghiên cứu đơn mạng, Sutoyo và cộng sự (2021) đã đề xuất mô hình sử dụng CNN để khám phá cảm xúc sinh viên về năng lực sư phạm giảng viên. Mô hình áp dụng trên các phản hồi của những câu hỏi mở của bảng câu hỏi EDOM. Kết quả thực nghiệm đạt Accuracy, Precision, Recall và F1-Score lần lượt là 87,95%, 87%, 78% và 81%. Bên cạnh đó, Onan (2020) cũng trình bày một mô hình dựa trên RNN để khai thác ý kiến đánh giá giáo viên nhằm đo lường hiệu quả giảng dạy và ra quyết định. Trong đó, RNN có cơ chế chú ý kết hợp biểu diễn dựa trên lược đồ nhúng từ GloVe. Thực nghiệm với 154.000 đánh giá cho kết quả Accuracy 98,29% ở tác vụ cực cảm xúc, vượt trội so với các phương pháp học máy thông thường. Cùng ý tưởng RNN nhưng ở dạng biến thể, Kandhro và cộng sự (2019) đã phát triển một mô hình sử dụng lớp nhúng từ đào tạo trước kết hợp LSTM, có khả năng thu thập thông tin ngữ nghĩa và cú pháp quan trọng. Mô hình cho thấy hiệu quả xác định cảm xúc trong các đánh giá giáo viên từ sinh viên và tiềm năng khắc phục một số hạn chế của phương pháp truyền thống. Ngoài ra, Sindhu và cộng sự (2019) để xuất một mô hình khai thác ý kiến ở cấp độ khía cạnh dựa trên LSTM hai lớp. Trong đó, lớp thứ nhất thực hiện dự đoán các khía cạnh và lớp thứ hai sẽ xác định cực cảm xúc của các khía cạnh được dự đoán. Thực nghiệm tiến hành trên dữ liệu đánh giá của Đại học Sukkur IBA, đạt Accuracy ở tác vụ trích xuất khía cạnh 91% và phân loại cực cảm xúc là 93%.

Trong khi xét các nghiên cứu đa mạng kết hợp, sự đa dạng được thể hiện trong xây dựng kiến trúc mô hình. Cụ thể, Kastrati và cộng sự (2020) đã xây dựng một bộ khung SA cấp độ khía cạnh cho các đánh giá khoá học trực tuyến đại chúng mở (MOOC). Bộ khung này sử dụng các chú thích nhãn giám sát yếu về các khía cạnh MOOC để giảm gắn nhãn thủ công. Đối với hai tác vụ chính, CNN được sử dụng để trích xuất khía cạnh và chuyển sang CNN-LSTM để thực hiện phân loại cực cảm xúc của khía cạnh. Thực nghiệm trên khoảng 105.000 đánh giá từ Coursera và 5.989 đánh giá từ sinh viên lớp học truyền thống cho thấy kết quả đầy khả quan. Để tăng cường hiệu quả của LSTM, Peng và cộng sự (2022) đề xuất một mô hình SA đánh giá giảng dạy, kết hợp CNN và BiLSTM nhằm nâng cao khả năng trích xuất thông tin. Bên cạnh đó, cơ chế chú ý cũng được sử dụng nhằm tìm kiếm sự liên kết giữa văn bản trong các đánh giá với cảm xúc. Thực nghiệm cho thấy hiệu quả của đề xuất với giá trị F1 tối thiểu 0,748, vượt trội các mô hình khác. Thay thế BiLSTM bằng GRU, Das và cộng sự (2022) đã thiết kế một mô hình SA cho nền tảng e-learning. Mô hình sử dụng lược đồ nhúng GloVe để thực hiện vector hoá dữ liệu văn bản đầu vào. CNN và GRU cũng được tích hợp tuần tự trong mô hình để hỗ trợ cho mục đích phân loại cực cảm xúc và đã mang lại những kết quả thực nghiệm đầy khả quan.

Xem xét các nghiên cứu sử dụng BERT, Zheng và cộng sự (2020) đã thiết kế mô hình BERT-BiGRU cho các đánh giá khoá học trực tuyến. Cụ thể, BERT được dùng với vai trò là một bộ mã hoá câu đầu vào và BiGRU nhận kết quả mã hoá để phân tích thông tin cảm xúc hỗ trợ cho phân loại. Ngoài ra, nhóm tác giả còn đề xuất một hệ thống tích hợp mô hình giúp tăng tính ứng dụng. Thực nghiệm tiến hành trên bộ dữ liệu của tác giả và so sánh với các phương pháp truyền thống. Kết quả mô hình đạt được hiệu suất cao nhất với Accuracy 98,82%. Bên cạnh, Dyulicheva và Bilashova (2021) cũng ứng dụng mô hình đào tạo trước BERT cho đánh giá người học Udemy. Các đánh giá được thu thập từ 300 khoá học và phân thành hai nhóm khía cạnh dựa trên từ vựng, gồm giáo viên và chương trình. BERT áp dụng trên từng nhóm để tìm hiểu thái độ người học đối với các khía cạnh. Kết quả cho thấy, thái độ tiêu cực về chương trình cao hơn khi so với giáo viên.

Tại Việt Nam những năm gần đây, sự xuất hiện của bộ dữ liệu UIT-VSFC đã thúc đẩy các nghiên cứu SA trong lĩnh vực giáo dục. Dựa trên bộ dữ liệu này, nhiều mô hình theo các hướng tiếp cận khác nhau được giới thiệu nhằm tối ưu hoá hiệu suất các nhiệm vụ, trong đó có học sâu. Tuy nhiên, hầu hết nghiên cứu chỉ tập trung vào nhiệm vụ phân loại cực cảm xúc. Vấn đề mất cân bằng dữ liệu hầu như không được đề cập và giải quyết ở các nghiên cứu, F1-Score (Macro) ít được sử dụng trong hoạt động đánh giá. Dưới đây sẽ là một số nghiên cứu SA theo hướng tiếp cận học sâu liên quan lĩnh vực giáo dục, sử dụng bộ dữ liệu UIT-VSFC.

Đầu tiên, Nguyen V. X. Phu và cộng sự (2019) đã thực hiện một so sánh giữa các mô hình SA phân loại truyền thống và học sâu trên bộ dữ liệu UIT-VSFC. Nghiên cứu này nhằm mục đích tìm ra mô hình hiệu quả nhất dựa trên các tiêu chí, hỗ trợ nâng cao chất lượng đào tạo. Kết quả thực nghiệm cho thấy sự vượt trội của các mô hình phân loại theo hướng học sâu. BiLSTM đạt hiệu suất cao nhất với F1-score (Micro) 92,0% ở nhiệm vụ phân loại cực cảm xúc và 89,6% ở nhiệm vụ phân loại chủ đề.

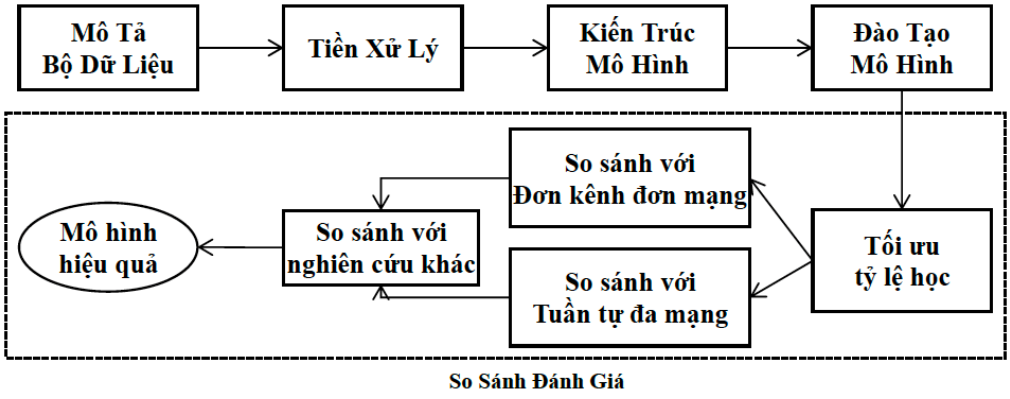
Xét cụ thể ở nhóm nghiên cứu đơn mạng, Nguyen Duc Vu và cộng sự (2018) đề xuất một hướng tiếp cận mới để xây dựng mô hình SA dựa trên Cây phụ thuộc và LSTM. Mô hình tập trung cho nhiệm vụ phân loại cực cảm xúc trong phản hồi học viên tại Việt Nam. Thực nghiệm cho thấy, hướng tiếp cận đề xuất đã mang lại kết quả vượt trội so với mô hình LSTM khi kết hợp Cây phụ thuộc, LSTM và bộ phân loại Support Vector Machine (SVM), với Accuracy 90,7% và F1-Score (Weighted) 90,2%. Nguyen Quan Hoang và cộng sự (2020) đã áp dụng hướng tiếp cận học sâu để đề xuất một mô hình SA với tên gọi ReAt-Bi-LSTM. Cụ thể, kỹ thuật residual sử dụng trong các lớp BiLSTM và cơ chế attention tích hợp sau lớp BiLSTM. Sau cùng, biểu diễn của dữ liệu đầu vào sẽ kết hợp giữa vector ngữ cảnh và đầu ra của BiLSTM. Kết quả đạt Accuracy 91,16% và F1-Score 90,42%.

Đối với nghiên cứu đa mạng kết hợp, Le Si Lac và cộng sự (2020) đưa ra một kiến trúc mới để khai thác các phản hồi của khách hàng và học viên tiếng Việt. Dựa trên ý tưởng tận dụng ưu điểm các mạng khác nhau, mô hình được cấu thành từ BiLSTM và CNN (sử dụng nhiều filters) theo tuần tự, nhằm trích lọc thông tin phục vụ xác định cực cảm xúc. Thực nghiệm chứng minh hiệu quả so với các dạng kết hợp khác giữa CNN, LSTM và BiLSTM, với F1-Score là 93,55%.

Trong khi đó, các nghiên cứu sử dụng BERT nhận được sự quan tâm lớn tại Việt Nam. Truong Trong Loc và cộng sự (2020) nghiên cứu sử dụng PhoBERT nhằm phát hiện và phân loại cảm xúc học viên để hỗ trợ Mô hình quy trình cải tiến liên tục. Các thông tin 4 lớp cuối của PhoBERT được nối lại làm đầu vào cho Multi-Layer Perceptron (MLP), với hàm Softmax phân loại. Kết quả thực nghiệm đạt Accuracy 94,28% và F1-Score (Weighted) 93,92%, cao hơn kết quả nghiên cứu của Nguyen Duc Vu và cộng sự (2018). Dựa trên ý tưởng mỗi một mô hình sẽ phù hợp với từng bộ dữ liệu cụ thể, Huynh Duc Huy và cộng sự (2020) đề xuất một mô hình hỗn hợp. Mô hình là sự kết hợp các mô hình đơn lẻ gồm BERT, CNN, LSTM và các biến thể. Nhóm đã thực nghiệm trên các bộ dữ liệu HSD-VLSP, UIT-VSMEC và UIT-VSFC. Trong đó, hiệu suất theo F1-score (Micro) đạt trên UIT-VSFC ở nhiệm vụ phân loại chủ đề là 89,70% và phân cực cảm xúc là 92,79%. Ngoài ra, Cu Vinh Loc và cộng sự (2022) đã đề cập tầm quan trọng của bình luận trực tuyến đối với các tổ chức, trong đó có bình luận học viên đối với nhà trường. Việc ứng dụng một mô hình SA tự động là một điều cần thiết. Nhóm tác giả sử dụng PhoBERT như một lớp nhúng, kết hợp CNN để ghi nhận thông tin cục bộ theo tuần tự. Thực nghiệm chứng minh được hiệu quả mô hình đề xuất với F1-score 91,43%, cao hơn so với các mô hình khác. Theo hướng tiếp cận lai, Dang N. Cach và cộng sự (2023) với mong muốn xây dựng một mô hình SA hiệu quả đã đề xuất kiến trúc kết hợp PhoBERT, CNN, LSTM và SVM. Với tuần tự lớp nhúng là PhoBERT, lớp trích xuất thông tin cục bộ là CNN, lớp trích xuất thông toàn chuỗi là LSTM và SVM tính toán xác suất phân loại nhãn. Mô

hình đạt Accuracy 93,52% và F1-Score 82,01%. Gần đây, Phan Chau Thang và cộng sự (2023) nhấn mạnh vấn đề mất cân bằng và nhiễu trong dữ liệu. Để giải quyết vấn đề trên, nhóm nghiên cứu đề xuất sử dụng cấu trúc biểu đồ của dữ liệu nhằm nắm bắt sự phụ thuộc về cú pháp và ngữ nghĩa trong các đánh giá thông qua Graph Convolutional Networks (GCN). Mô hình là sự kết hợp giữa GCN và PhoBERT, được gọi là ViCGCN. Thực nghiệm trên nhiều bộ dữ liệu đã cho thấy sự ưu việt của mô hình so với 13 mô hình khác trong việc giải quyết các vấn đề đã nêu. Tại bộ dữ liệu UIT-VSFC, ViCGCN (base) đạt F1-Score (Weighted) 94,12% và F1-Score (Macro) 83,67% ở nhiệm vụ phân loại cực cảm xúc, đạt F1-Score (Weighted) 90,12% và F1-Score (Macro) 80,11% ở nhiệm vụ phân loại chủ đề.

3. Phương pháp nghiên cứu



Hình 1. Quy trình nghiên cứu tổng quan

Nghiên cứu này tập trung xây dựng một mô hình SA phản hồi học viên tiếng Việt, có khả năng giải quyết các hạn chế hiệu quả hơn so với các nghiên cứu trước đây. Tổng quan toàn bộ quy trình nghiên cứu được thể hiện trong Hình 1. Đầu tiên, bộ dữ liệu nghiên cứu được xác định và mô tả về quy mô, số lượng nhãn, tỷ lệ nhãn,… hỗ trợ cho quá trình thiết kế, đào tạo và đánh giá. Sau đó, giai đoạn tiền xử lý dữ liệu sẽ xác định các bước làm sạch và chuyển đổi bộ dữ liệu để loại bỏ thông tin không giá trị và thay đổi định dạng phù hợp cho các mô hình. Giai đoạn tiếp theo, kiến trúc mô hình đề xuất sẽ được thiết kế chi tiết tại từng thành phần nhằm giải quyết các vấn đề. Từ đó, hoạt động đào tạo mô hình với dữ liệu tiền xử lý sẽ triển khai bằng những thiết lập siêu tham số và kịch bản cụ thể. Sau cùng, một chuỗi tác vụ được thực hiện trên mô hình nhằm tìm kiếm tỷ lệ học tối ưu, đánh giá so sánh với các kiến trúc mô hình khác và so sánh với nghiên cứu gần nhất để xem xét mức độ cải thiện hiệu quả của nghiên cứu này.

3.1. Mô tả bộ dữ liệu

Bộ dữ liệu UIT-VSFC được sử dụng cho hoạt động đào tạo và đánh giá trong nghiên cứu này. Với hơn 16.000 phản hồi tiếng Việt của sinh viên tại cơ sở đào tạo Đại học, UIT-VSFC được gán nhãn thủ công nhằm đảm bảo độ chính xác và hỗ trợ cho hai nhiệm vụ là phân loại chủ đề và cực cảm xúc. Vì vậy, UIT-VSFC sẽ có hai nhóm nhãn riêng biệt. Trong đó, nhiệm vụ phân loại chủ đề có bốn nhãn, bao gồm: giảng viên (lecturer), chương trình (curriculum), cơ sở (facility) và khác (others). Còn tại nhiệm vụ phân loại cực cảm xúc, ba nhãn được sử dụng, bao gồm: tích cực (positive), tiêu cực (negative) và trung tính (neutral). Bộ dữ liệu cũng được phân tách thành ba bộ riêng lẻ là train, dev và test phục vụ cho hoạt động đào tạo và đánh giá các mô hình học sâu theo tỷ lệ lần lượt 70,0%, 10,0%, và 20,0%.

Xét các nhãn cực cảm xúc, nhãn tích cực được gán cho các phản hồi mà sinh viên thể hiện

sự hài lòng về các khía cạnh trong quá trình học tập, ví dụ "giảng viên tận tâm, nhiệt huyết trong giảng dạy" được gán nhãn là tích cực. Ngược lại, nhãn tiêu cực được gán cho các phản hồi mà sinh viên thể hiện sự không hài lòng về các khía cạnh, ví dụ "giảng bài kém thu hút, vị trí đứng giảng không hợp lý, ôn tập cuối kỳ còn chưa tập trung vào vấn đề" là một phản hồi tiêu cực. Còn đối với trung tính, nhãn này được gán cho những phản hồi không hoàn chỉnh, không rõ ràng cảm xúc hoặc không chứa ý kiến của sinh viên, ví dụ: "Em cảm ơn thầy" là một phản hồi được gán nhãn trung tính do không chứa các từ thể hiện cảm xúc.

Xét các nhãn chủ đề, nhãn giảng viên được gán cho các phản hồi thể hiện cảm xúc về các mặt liên quan đến giảng viên như phương pháp giảng dạy, thái độ hoặc trình độ,… Ví dụ "giảng viên giải thích kỹ và chi tiết". Nhãn chương trình được gán cho các phản hồi liên quan đến chương trình như bài tập, các bài lab, nội dung giảng dạy, kiến thức,... Ví dụ: "môn học này giúp chúng em hiểu ra những vấn đề cơ bản". Nhãn cơ sở được gán cho các phản hồi liên quan đến cơ sở vật chất như máy vi tính, máy chiếu, máy lạnh, hệ thống chiếu sáng,... Ví dụ: "máy chiếu nhiều lúc chẳng muốn để nhìn, chất lượng kém, ánh sáng làm mờ.". Ngoài ra, các phản hồi không thuộc các nhãn trên hoặc không rõ ràng sẽ được gán nhãn là khác, ví dụ: "cám ơn đã dạy lớp em.". Một số ví dụ cho phản hồi được gán cả nhãn chủ đề và cực cảm xúc. Phản hồi "nhiệt tình giảng dạy, gần gũi với sinh viên" thể hiện sự hài lòng đối với giảng viên nên được gán nhãn "giảng viên" và nhãn "tích cực". Phản hồi "slide giáo trình đầy đủ" thể hiện sự hài lòng đối với tài liệu thuộc chương trình nên được gán nhãn "chương trình" và nhãn "tích cực". Phản hồi "thời lượng học quá dài, không đảm bảo tiếp thu hiệu quả" thể hiện sự không hài lòng đối với thời lượng chương trình nên được gán nhãn "chương trình" và nhãn "tiêu cực".

Phân tích chi tiết tỷ lệ các nhãn như hình 2, một sự mất cân bằng tồn tại trong từng nhóm được thể hiện. Trong nhóm nhãn chủ đề, nhãn giảng viên chiếm đa số với 71,76%, phần còn lại thuộc về nhãn chương trình 18,79%, nhãn cơ sở và khác chiếm tỷ lệ còn lại lần lượt chỉ 4,4% và 5,04%. Còn tại nhóm nhãn cực cảm xúc, nhãn tích cực và tiêu cực là hai nhãn chiếm đa số với tỷ lệ tương đối đồng đều là 49,69% và 45,99%, phần còn lại thuộc về nhãn trung tính với chỉ 4,32%. Điều này có khả năng gây nên hiệu suất phân loại kém ở các nhãn tỷ lệ thấp trong nhóm, làm giảm độ chính xác trung bình. Đây sẽ là vấn đề cần giải quyết và cải thiện trong nghiên cứu này.

3.2. Tiền xử lý

Mục đích chính của tiền xử lý là chuẩn bị dữ liệu đầu vào cho các tác vụ nối tiếp. Tiền xử lý bao gồm một chuỗi các bước làm sạch dữ liệu như: Xoá dấu câu; Chuẩn hoá từ ngữ; Phân đoạn từ (segmentation); Xoá các ký tự không cần thiết. Chi tiết như sau:

- Xóa dấu câu: Loại bỏ các dấu câu như dấu chấm, dấu phẩy, dấu hỏi,… trong các câu. Đây là những thành phần không mang lại giá trị phân loại cho đầu vào.
- Chuẩn hóa từ ngữ: Các từ viết hoa trong câu được chuyển đổi thành từ viết thường. Sau đó, các từ viết tắt hay sai chính tả sẽ được thay thế bằng các từ đúng chuẩn trong tiếng Việt. Đồng thời, các ký hiệu emoji cũng sẽ chuyển đổi thành các từ ngữ tương ứng.
- Phân đoạn từ (segmentation): Xác định các từ ghép trong tiếng Việt và gom nhóm lại thông qua việc chèn "_" ở giữa hai từ đơn thuộc một từ ghép. Thư viện VnCoreNLP (Vu Thanh và cộng sự, 2018) sẽ là lựa chọn để thực hiện ở bước này.
- Xóa ký tự không cần thiết: Kiểm tra và loại bỏ các ký tự không cần thiết, không phải là từ ngữ. Trong bước này, các khoảng trống liên tục cũng sẽ điều chỉnh thành một khoảng trống để giảm bớt độ dài của câu.

Một điểm đáng lưu ý, nghiên cứu không thực hiện xoá các từ dừng (stop words). Nguyên nhân do hoạt động này sẽ vô tình xoá đi các thông tin quan trọng, làm ảnh hưởng đến kết quả phân loại.

Bên cạnh làm sạch dữ liệu, hoạt động chuyển đổi dữ liệu thành định dạng đầu vào phù hợp sẽ tiếp nối trong giai đoạn tiền xử lý. Cụ thể, các câu được phân tách (tokenizing) theo khoảng trống thành những từ đơn lẻ và mã hoá thành các mã số đại diện dựa trên thư viện từ vựng PhoBERT. Sau đó, hoạt động chuyển đổi định dạng sẽ áp dụng trên mỗi câu thành input ids, token type ids và attention mask theo yêu cầu đầu vào PhoBERT

3.3. Kiến trúc mô hình đề xuất

Sau giai đoạn tiền xử lý, hoạt động thiết kế mô hình được tiến hành với kiến trúc hai thành phần chính. Chi tiết về các thành phần theo tuần tự xử lý dữ liệu như sau.

Thành phần thứ nhất có chức năng biểu diễn đặc trưng, tạo các vector biểu diễn bằng cách mã hóa thông tin các câu đầu vào. Nghiên cứu này sử dụng PhoBERT, một mô hình đào tạo trước và tinh chỉnh chuyên biệt cho tiếng Việt. PhoBERT dựa trên mô hình BERT phổ biến và thường sử dụng để tạo ra các biểu diễn văn bản. Đầu ra của thành phần này sẽ nhận bốn trạng thái ẩn (hidden states) cuối cùng của PhoBERT và nối lại với nhau, trước khi chuyển sang thành phần tiếp.

Thành phần thứ hai là một kiến trúc đa kênh, bao gồm hai kênh là CNN và GRU đặt song song, nhận đồng thời đầu vào là đầu ra của thành phần thứ nhất. Đây là thành phần có vai trò trích xuất thông tin dựa trên đặc điểm từng mạng. Theo đó, CNN có ưu điểm về trích xuất thông tin sự phụ thuộc cục bộ giữa các dữ liệu gần nhau nhưng khó nắm bắt được sự phụ thuộc xa. Trong khi, GRU là biến thể RNN mang kiến trúc đơn giản hơn LSTM nhưng vẫn đảm bảo hiệu quả (Chung và cộng sự, 2014), có ưu điểm nắm bắt thông tin tổng thể toàn chuỗi hay sự phụ thuộc xa nhưng lại hạn chế khi trích xuất sự phụ thuộc gần. Bằng cách sử dụng hai mạng cho hai kênh, nghiên cứu này kỳ vọng rằng ưu và nhược điểm của từng mạng sẽ bù trừ và hỗ trợ lẫn nhau, từ đó cải thiện hiệu quả trích xuất thông tin. Cụ thể về đặc điểm thiết kế của từng kênh như sau:

- Kênh CNN: Mạng sẽ bao gồm 3 lớp tính chập 1 chiều (Conv1D) song song. Mỗi lớp tính chập có 150 bộ lọc (Filter) với kích thước của các bộ lọc lần lượt là 3, 5, 7. Sau đó, các bản đồ đặc trưng (Feature map) là đầu ra của các lớp tính chập được đưa qua các lớp gộp tối đa 1 chiều (MaxPool1D) nhằm rút trích thông tin quan trọng hay giảm chiều kết quả. Sau cùng, những thông tin rút trích sẽ kết hợp lại để đưa qua lớp làm phẳng (Flatten) tạo thành đầu ra của kênh CNN.

- Kênh GRU: Để nâng cao hiệu quả trích xuất thông tin đa chiều, Bi-GRU được lựa chọn sử dụng. Theo đó, hai GRU sẽ nối tiếp và đặt ngược chiều nhau. Điều này giúp cho việc nắm bắt thông tin tổng thể của chuỗi theo cả hai hướng. Về số lượng đơn vị (units) trong mỗi GRU, nghiên cứu này sẽ lựa chọn 128, đây là số lượng phổ biến trong RNN và các biến thể. Cuối cùng, đầu ra của Bi-GRU sẽ là đầu ra của kênh GRU.

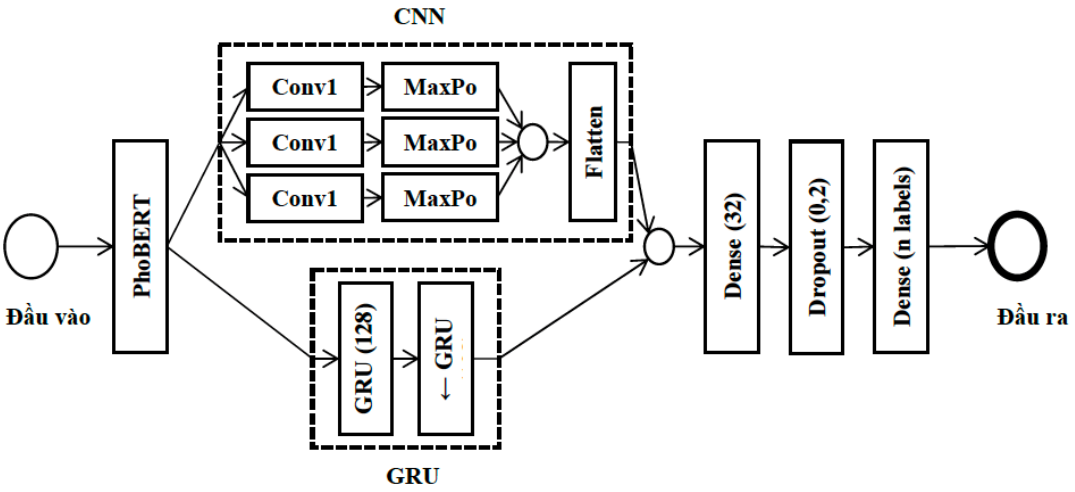
Kết quả của hai kênh sẽ được kết hợp tạo đầu ra của thành phần thứ hai và đưa vào một số lớp hỗ trợ phân loại cuối cùng. Cụ thể, các lớp được đặt lần lượt như sau:

- Lớp dày đặc thứ nhất (Dense): Lớp này sẽ nhận trực tiếp đầu ra từ thành phần thứ hai để hỗ trợ phân tích thông tin trích xuất. Cấu tạo của lớp này sẽ bao gồm 32 nơ-ron sử dụng hàm ReLu.

- Lớp từ bỏ (Dropout): Lớp này được sử dụng để tránh trường hợp quá khớp (Overfitting) khi khiến trúc mô hình có nhiều lớp và nhiều trọng số. Cấu hình dành cho lớp này sẽ 0,2.

- Lớp dày đặc thứ hai (Dense): Để nhận kết quả phân loại cuối cùng, một lớp dày đặc sẽ được đặt ở cuối kiến trúc mô hình. Trong đó, số lượng nơ-ron cho lớp này sẽ bằng đúng số lượng nhãn cần phân loại, và sử dụng hàm softmax cho bài toán phân loại đa lớp.

Hình 2 mô tả chi tiết về kiến trúc mô hình đề xuất của nghiên cứu.



Hình 2. Kiến trúc mô hình đề xuất

3.4. Đào tạo mô hình

Để tiến hành đào tạo mô hình đề xuất, một số thiết lập được xác định. Đầu tiên, nghiên cứu này xác định đào tạo hai mô hình dựa trên kiến trúc mô hình đề xuất cho hai nhiệm vụ riêng biệt là chủ đề và cực cảm xúc. Với số lượng nhãn cần phân loại của nhiệm vụ chủ đề là bốn nên lớp dày đặc thứ hai sẽ có bốn nơ-ron. Tại nhiệm vụ cực cảm xúc, ba nhãn cần phân loại nên số lượng nơ-ron trong lớp dày đặc thứ hai sẽ là ba.

Dữ liệu đầu vào sẽ chia theo batch size là 32 để đưa vào đào tạo mô hình với epoch là 10 tương ứng 10 lần lặp đào tạo bộ dữ liệu. Trình tối ưu sử dụng sẽ là Adam với số bước khởi động (num warmup steps) là 10% số bước đào tạo ban đầu nhằm tối thiểu mất mát và tìm kiếm trọng số mô hình. Ngoài ra, nghiên cứu này còn bổ sung một trình giám sát đào tạo, theo dõi chỉ số mất mát xác thực (val loss), nếu không có sự cải thiện sau 2 epoch thì mô hình sẽ dừng đào tạo để tránh quá khớp. Đối với tỷ lệ học (learning rate), siêu tham số này sẽ xác định giá trị tối ưu ở giai đoạn so sánh đánh giá.

3.5. So sánh đánh giá

Với mục đích so sánh và đánh giá mô hình đề xuất, cơ sở so sánh sẽ dựa trên các chỉ số đánh giá bài toán phân loại. Cụ thể, 3 chỉ số gồm F1-Score (Macro), F1-Score (Weighted) và Accuracy được dùng để đo lường hiệu suất mô hình, đây cũng chính các cơ sở để lựa chọn mô hình phù hợp. Theo đó, F1-Score là một thước đo quan trọng để đánh giá hiệu suất của mô hình phân loại ở từng nhãn, phản ánh sự cân bằng giữa Precision xác định chất lượng các dự đoán của mô hình và Recall đo lường khả năng phát hiện chính xác hay bỏ sót của mô hình. Tại mỗi nhãn, F1-Score cao biểu thị sự cân bằng tốt, thể hiện mô hình đạt được Precision và Recall đều cao, ngược lại với F1-Score thấp thường biểu thị sự thiếu cân bằng hay có sự đánh đổi giữa Precision và Recall trong mô hình. Để đánh giá tổng thể mô hình, F1-Score của các nhãn sẽ được tổng hợp để xác định F1-Score (Macro) bằng trung bình cộng và F1-Score (Weighted) bằng trung bình có trọng số với trọng số theo số lượng xuất hiện mỗi nhãn. Trong khi với Accuracy, đây cũng là một thước đo phổ biến trong bài toán phân loại, đo lường tỷ lệ dự đoán đúng so với tổng số dự đoán được thực hiện.

Đầu tiên, việc xác định tỷ lệ học (learning rate) tối ưu sẽ là công việc mà nghiên cứu này ưu tiên thực hiện. Tỷ lệ học là một trong những siêu tham số quan trọng, có ảnh hưởng lớn đến hiệu suất mô hình. Do vậy, một loạt thử nghiệm đào tạo và đánh giá mô hình sẽ triển khai trên 5 mức tỷ lệ học khác nhau ở từng nhiệm vụ. Hoạt

động phân tích sẽ triển khai trên các kết quả, nhằm xác định tỷ lệ học tối ưu thông qua các chỉ số đánh giá.

Dựa trên các tỷ lệ học tối ưu đã xác định, các hoạt động thực nghiệm đánh giá liên quan đến kiến trúc ở giai đoạn sau được thực hiện. Trong đó, nghiên cứu sẽ đánh giá hiệu quả trong việc ứng dụng kiến trúc đa kênh đa mạng so sánh với kiến trúc đơn kênh đơn mạng. Để thực hiện, kiến trúc mô hình đề xuất sẽ điều chỉnh cắt giảm kênh, tạo thành hai mô hình đơn kênh sử dụng CNN và GRU riêng lẻ. Ngoài ra, kiến trúc đa kênh đa mạng cũng sẽ so sánh với kiến trúc tuần tự đa mạng thường được sử dụng trong các nghiên cứu trên thế giới và tại Việt Nam. Theo đó, mỗi kiến trúc so sánh sẽ sử dụng cùng lúc CNN và GRU nhưng có sự đảo vị trí lần lượt để tạo ra các biến thể là CNN-GRU và GRU-CNN.

Cuối cùng, nghiên cứu này sẽ xem xét kết quả đạt được của mô hình đề xuất so với mô hình ViCGCN trên bộ dữ liệu UIT-VSFC, đã đề cập trong các nghiên cứu liên quan. Đây là mô hình có cùng ý tưởng nghiên cứu, về việc giải quyết vấn đề mất cân bằng dữ liệu ở cả hai nhiệm vụ. Hai chỉ số F1-Score (Macro) và F1-Score (Weighted) sẽ là cơ sở để xem xét. Điều này mang lại một góc nhìn khách quan giữa mô hình đề xuất so với nghiên cứu trước đây, giúp xác định ưu và nhược điểm thực sự của mô hình.

4. Kết quả và thảo luận

Về xác định tỷ lệ học tối ưu, kết quả từ thực nghiệm tại bảng 1 cho thấy ở từng nhiệm vụ sẽ nhận một tỷ lệ học có hầu hết chỉ số vượt trội so với phần còn lại. Chỉ số F1-Score (Weighted) và Accuracy không có sự chênh lệch quá lớn khi tỷ lệ học thay đổi ở từng nhiệm vụ, lần lượt không quá 0,015 và 0,012. Trái ngược, F1-Score (Macro) lại có sự chênh lệch đáng kể khi thay đổi tỷ lệ học, gần 0,037 tại nhiệm vụ chủ đề và gần 0,024 tại nhiệm vụ cực cảm xúc.

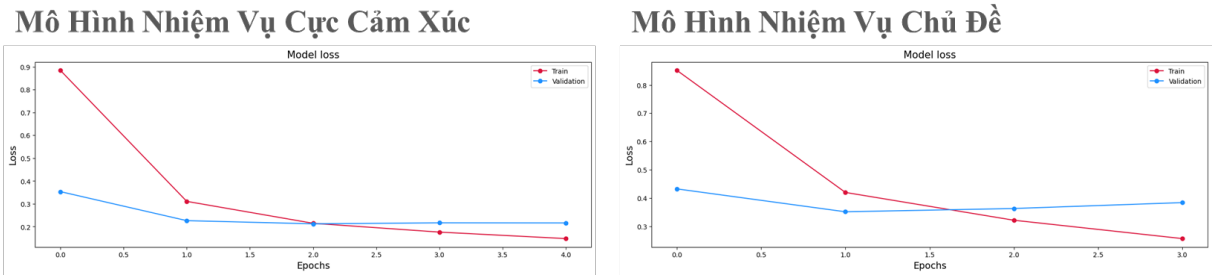
Bảng 1. Hiệu suất mô hình đề xuất theo tỷ lệ học

Nhiệm Vụ	Tỷ lệ học	F1-Score (Macro)	F1-Score (Weighted)	Accuracy
Chủ đề	0,00005	0,8008	0,8889	0,8913
	0,00004	0,8111	0,8961	0,8952
	0,00003	0,7744	0,8819	0,8838
	0,00002	0,7924	0,8895	0,8907
	0,00001	0,8011	0,8913	0,8929
Cực cảm xúc	0,00005	0,8273	0,9309	0,9327
	0,00004	0,8421	0,9361	0,9368
	0,00003	0,8189	0,9337	0,9378
	0,00002	0,8418	0,9384	0,9413
	0,00001	0,8333	0,9374	0,9406

Xét cụ thể tại nhiệm vụ chủ đề, tỷ lệ học 0,00004 cho thấy sự vượt trội ở các chỉ số so với các tỷ lệ học khác, với F1-Score (Macro) 0,8111, F1-Score (Weighted) 0,8961 và Accuracy 0,8952. Thể hiện này giúp dễ dàng xác định tỷ lệ học tối ưu tại nhiệm vụ chủ đề là 0,00004. Trong khi tại nhiệm vụ cực cảm xúc, 0,00004 và 0,00002 là hai tỷ học có sự vượt trội so với phần còn ở một vài chỉ số riêng lẻ. F1-Score (Macro) đạt cao nhất là 0,8421 tại tỷ lệ học 0,00004. F1-Score (Weighted) cùng với Accuracy đạt cao nhất tại tỷ lệ học 0,00002 với giá trị lần lượt là 0,9384 và 0,9413. Xét chi tiết về sự chênh lệch, tỷ lệ học 0,00004 tuy đạt F1-Score (Macro) cao nhất nhưng không

đáng kể so với tỷ lệ học 0,00002, cao hơn chỉ 0,0003 nhưng F1-Score (Weighted) và Accuracy lại suy giảm lần lượt 0,0023 và 0,0045. Do đó, nghiên cứu này chọn 0,00002 là tỷ lệ học tối ưu tại nhiệm vụ cực cảm xúc.

Với hai tỷ lệ học tối ưu cho hai nhiệm vụ, các kết quả xác thực tại mỗi epoch trong quy trình đào tạo được xem xét như trong hình 3. Theo đó, cả hai quy trình đào tạo đều có sự suy giảm mất mát xác thực (validation loss) và mất mát đào tạo (train loss) đáng kể tại epoch thứ hai. Sau epoch này, mô hình tại nhiệm vụ cực cảm xúc có sự suy giảm nhẹ mất mát xác thực tại epoch thứ ba nhưng tăng dần ở hai epoch còn lại. Trong khi, mô hình tại nhiệm vụ chủ đề lại có sự tăng dần mất mát xác thực tại epoch thứ ba và thứ tư. Tất cả cho thấy các mô hình đã sớm tối ưu ngay trong khoảng nửa lịch trình đào tạo 10 epoch. Cụ thể, mô hình nhiệm vụ cực cảm xúc tối thiểu tại epoch thứ ba và nhiệm vụ chủ đề tối thiểu tại epoch thứ hai. Đây là kết quả của các thiết lập đào tạo, góp phần cắt giảm thời gian đào tạo mô hình và tiết kiệm chi phí đáng kể so với kế hoạch.



Hình 3. Quy trình đào tạo của các mô hình

Về đánh giá hiệu quả ứng dụng kiến trúc đa kênh đa mạng, việc sử dụng đồng thời hai kênh kết hợp là CNN và GRU đã cho thấy sự cải thiện hiệu suất của mô hình đề xuất khi so sánh với các mô hình đơn kênh đơn mạng. Bảng 2 trình bày hiệu suất các mô hình này. Cụ thể, các chỉ số đánh giá của mô hình mô hình đề xuất PhoBERT-Multi(CNN-GRU) đều đạt sự vượt trội ở các nhiệm vụ so với các mô hình khác. Theo đó, mô hình đề xuất tại nhiệm vụ chủ đề cao hơn các mô hình khác với F1-Score (Macro) ít nhất 0,0180, F1-Score (Weighted) ít nhất 0,0075 và Accuracy ít nhất 0,0042. Còn tại nhiệm vụ cực cảm xúc, mô hình đề xuất cao hơn các mô hình khác với F1-Score (Macro) ít nhất 0,0134, F1-Score (Weighted) ít nhất 0,0072 và Accuracy ít nhất 0,0019. Điều này cho thấy việc tăng cường thêm mạng, tăng khối lượng tính toán để trích xuất thông tin đã cải thiện hiệu suất. Đặc biệt, F1-Score (Macro) là chỉ số có mức cải thiện đáng kể nhất khi kiến trúc đa kênh đa mạng được ứng dụng.

Bảng 2. So sánh hiệu suất giữa kiến trúc đa kênh đa mạng và đơn kênh đơn mạng

Nhiệm Vụ	Models	F1-Score (Macro)	F1-Score (Weighted)	Accuracy
Chủ đề	PhoBERT-GRU	0,7855	0,8849	0,8907
	PhoBERT-CNN	0,7931	0,8886	0,8910
	PhoBERT-Multi(CNN-GRU)	0,8111	0,8961	0,8952
Cực cảm xúc	PhoBERT-GRU	0,8181	0,9312	0,9337
	PhoBERT-CNN	0,8284	0,8886	0,9394
	PhoBERT-Multi(CNN-GRU)	0,8418	0,9384	0,9413

Về đánh giá hiệu quả kiến trúc đa kênh thuộc mô hình đề xuất so với kiến trúc tuần tự, hiệu suất thể hiện trong Bảng 3 đã chứng minh tính vượt trội của mô hình đề xuất so với phần còn lại. Xét tại nhiệm vụ chủ đề, kiến trúc đa kênh vượt kiến trúc tuần tự tại F1-Score (Macro) ít nhất 0,0215, F1-Score (Weighted) ít nhất 0,0101 và Accuracy ít nhất 0,0089. Còn ở nhiệm vụ cực cảm xúc, kiến trúc đa kênh vượt kiến trúc tuần tự tại F1-Score (Macro) ít nhất 0,0191, F1-Score (Weighted) ít nhất 0,0032 và Accuracy ít nhất 0,0016. Nhìn chung, kiến trúc tuần tự được sử dụng phổ biến trong những nghiên cứu trước đây nhưng chưa phù hợp cho mục tiêu chống mất cân bằng dữ liệu. Các mô hình theo kiến trúc này có sự thua kém đáng kể tại F1-Score (Macro) so với kiến trúc đa kênh, sự chênh lệch có thể lên đến 0,4077 tại nhiệm vụ chủ đề và 0,2110 tại nhiệm vụ cực cảm xúc. Điều này tiếp tục củng cố sự hiệu quả của kiến trúc đa kênh đa mạng về khả năng chống mất căn bằng.

Bảng 3. So sánh hiệu suất giữa kiến trúc đa kênh đa mạng và tuần tự đa mạng

Nhiệm Vụ	Models	F1-Score (Macro)	F1-Score (Weighted)	Accuracy
Chủ đề	PhoBERT-CNN-GRU	0,7896	0,8860	0,8863
	PhoBERT-GRU-CNN	0,4034	0,8027	0,8339
	PhoBERT-Multi(CNN-GRU)	0,8111	0,8961	0,8952
Cực cảm xúc	PhoBERT-CNN-GRU	0,8227	0,9352	0,9397
	PhoBERT-GRU-CNN	0,6308	0,8956	0,9204
	PhoBERT-Multi(CNN-GRU)	0,8418	0,9384	0,9413

So sánh với mô hình ViCGCN, mô hình đề xuất PhoBERT-Multi(CNN-GRU) đã thể hiện sự cải thiện trong giải quyết vấn đề mất cân bằng trên bộ dữ liệu UIT-VSFC. Theo bảng 4, F1-Score (Macro) của mô hình đề xuất đã có sự gia tăng đáng kể khi ở nhiệm vụ chủ đề tăng 0,01 và nhiệm vụ cực cảm xúc tăng 0,0051. Tuy nhiên, F1-Score (Weighted) của mô hình đề xuất lại bị sụt giảm nhẹ khi nhiệm vụ chủ đề giảm 0,0051 và nhiệm vụ cực cảm xúc giảm 0,0028. Nhưng nhìn chung, sự sụt giảm F1-Score (Weighted) là không quá đáng kể khi so với sự gia tăng F1-Score (Macro) đã đạt được. Đây là một sự đánh đổi để đạt được một kết quả đáng ghi nhận và đạt được mục tiêu nghiên cứu này.

Bảng 4. So sánh hiệu suất giữa mô hình đề xuất và ViCGCN

Nhiệm Vụ	Models	F1-Score (Macro)	F1-Score (Weighted)
Chủ đề	ViCGCN	0,8011	0,9012
	PhoBERT-Multi(CNN-GRU)	0,8111	0,8961
Biến động		+0,0100	-0,0051
Cực cảm xúc	ViCGCN	0,8367	0,9412
	PhoBERT-Multi(CNN-GRU)	0,8418	0,9384
Biến động		+0,0051	-0,0028

Xem xét hiệu suất phân loại của các nhãn trong bảng 5. Sự chênh lệch hiệu suất đo lường bằng F1-Score giữa các nhãn trong nhóm vẫn thể hiện. Cụ thể, nhãn Khác trong nhóm chủ đề đạt F1-Score thấp nhất chỉ 0,5773, chênh lệch nhãn đạt hiệu suất cao hơn liền kề là Chương trình 0,2107. Còn tại nhóm cực cảm xúc, Trung tính là nhãn đạt F1-Score thấp nhất chỉ 0,6121, chênh lệch nhãn Tích cực đạt hiệu suất cao hơn liền kề lên đến 0,3428. Điều này cho thấy,

mô hình đề xuất mặc dù đã cải thiện khả năng chống mất cân bằng so với nghiên cứu trước đây là mô hình ViCGCN của Phan Chau Thang và cộng sự (2023), tác động của mất cân bằng dữ liệu được giảm đi đáng kể nhưng vẫn còn tồn tại.

Đánh giá cụ thể tại nhãn Khác và Trung tính, Precision và Recall của các nhãn này đều đạt mức thấp nhất trong nhóm với chỉ số lần lượt của Khác là 0,6364 và 0,5283, Trung tính là 0,7544 và 0,5150. Đây cũng chính là nguyên nhân dẫn đến F1-Score thấp. Theo đó, Precision thấp cho thấy, mô hình hạn chế khả năng tìm đúng tại các nhãn này, đặc biệt là nhãn Khác chỉ 0,6364. Còn đối với Recall thấp gần bằng 0,5000, vấn đề này cũng sẽ dẫn đến việc thường xuyên phân loại bỏ sót tại hai nhãn. Đây sẽ là một điều cần lưu ý đối với các nhà quản lý cơ sở giáo dục tại hai nhãn này khi sử dụng mô hình trong thực tế. Tuy nhiên, nhãn Khác và Trung tính là những nhãn được xây dựng nhằm mục đích dành cho các trường hợp phản hồi ngoại lệ hay không thuộc các nhãn còn lại trong mỗi nhóm. Những trường hợp này thường không phổ biến nên chiếm tỷ lệ thấp trong bộ dữ liệu và cũng không mang lại quá nhiều ý nghĩa khi phân loại. Do vậy, hiệu suất phân loại tại các nhãn này chưa cao sẽ không ảnh hưởng đáng kể đến tính ứng dụng của mô hình so với các các nhãn còn lại.

Bảng 5. Hiệu suất phân loại của mô hình đề xuất theo các nhóm nhãn

Nhiệm Vụ	Nhãn	Precision	Recall	F1-Score	Support
Chủ đề	Giảng viên	0,9411	0,9415	0,9413	2290
	Chương trình	0,7709	0,8059	0,7880	572
	Cơ sở	0,9379	0,9379	0,9379	145
	Khác	0,6364	0,5283	0,5773	159
Cực cảm xúc	Tích cực	0,9502	0,9597	0,9549	1590
	Tiêu cực	0,9461	0,9709	0,9583	1409
	Trung tính	0,7544	0,5150	0,6121	167

Dựa vào những chỉ số trên, nghiên cứu cho thấy, mô hình đề xuất có sự hiệu quả trong việc sử dụng đồng thời 3 mạng gồm PhoBERT, GRU, CNN và cách sắp xếp đa kênh cho GRU và CNN, so với các nghiên cứu liên quan. Tất cả đã được chứng minh thông qua đánh giá hiệu suất, so sánh với các lựa chọn đơn kênh đơn mạng, tuần tự với đa mạng mà các nghiên cứu trước đây sử dụng, cùng với đó là so sánh hiệu suất với mô hình ViCGCN có liên quan chặt chẽ và gần nhất. Sự hiệu quả này đến từ việc kế thừa các nghiên cứu liên quan trong việc sử dụng BERT làm lớp nhúng, sử dụng CNN và các mạng biến thể của RNN (LSTM, GRU). Trong đó, GRU là sự kế thừa từ các nghiên cứu ngoài nước để áp dụng vào mô hình đề xuất. Với những kết quả trên, mô hình đề xuất có khả năng phân loại cực và chủ đề của cảm xúc, được xây dựng dựa trên UIT-VSFC có hiệu suất cải thiện so với nghiên cứu gần nhất ở F1-Score (Macro) đã đáp ứng các mục tiêu đặt ra của nghiên cứu này.

Từ mục tiêu xây dựng mô hình, nghiên cứu này chỉ tập trung vào việc thiết kế, thực nghiệm, so sánh và đánh giá hiệu suất của mô hình đề xuất. Về nghiên cứu khoa học, kết quả nghiên cứu này đã đóng góp một kiến trúc mô hình mới trong lĩnh vực nghiên cứu SA trên phản hồi học viên tiếng Việt, thúc đẩy nghiên cứu thông qua việc sử dụng các bộ dữ liệu giáo dục sẵn có khác và tạo cơ sở để ứng dụng vào các nghiên cứu mở rộng. Ngoài ra, nghiên cứu này còn đóng góp một ý tưởng mới trong việc giải quyết bài toán mất cân bằng dữ liệu thường gặp trong lĩnh vực nghiên cứu về máy học có giám sát, hỗ trợ cho các hoạt động xây dựng mô hình

thuộc các miền khác được diễn ra hiệu quả. Về ứng dụng thực tiễn, mô hình thuộc lĩnh vực nghiên cứu là giáo dục và được xây dựng trên bộ dữ liệu là phản hồi học viên, do vậy mà kết quả của nghiên cứu có thể được ứng dụng trong các cơ sở giáo dục để giải quyết các bài toán liên quan đến quản lý giáo dục, dành cho các nhà quản lý. Một số bài toán có thể kể đến như:

- *Cải tiến chất lượng giảng dạy:* Bằng khả năng phân loại của mô hình nghiên cứu, nhà quản lý cơ sở giáo dục có thể hiểu rõ quan điểm trong phản hồi học viên. Các quan điểm sẽ được tự động xác định chủ đề với cực cảm xúc thể hiện một cách nhanh chóng, chính xác, và được tổng hợp, phân chia theo các chủ đề cụ thể, như giảng viên, chương trình, cơ sở,… để mang lại cho nhà quản lý một đánh giá tổng quan ở từng khía cạnh của cơ sở giáo dục được học viên đề cập với cảm xúc đi kèm. Điều này tạo điều kiện cho việc đánh giá các khía cạnh của cơ sở một cách toàn diện. Từ đó, các vấn đề là các khía cạnh với tỷ lệ tiêu cực cao sẽ xuất hiện, nhà quản lý có thể dễ dàng nhận ra những vấn đề cụ thể đang tồn tại và giải quyết bằng những hành động phù hợp, kịp thời, dựa trên dữ liệu cụ thể thay vì chỉ dựa trên cảm nhận hoặc phỏng đoán chủ quan. Với việc mô hình xử lý tốt vấn đề mất cân bằng dữ liệu, đã đảm bảo những vấn đề ít xuất hiện cũng được xử lý chính xác hơn, tránh tình trạng bỏ sót. Điều này mang lại sự toàn diện trong việc cải tiến chất lượng.

- *Điểm chuẩn (Benchmarking) các cơ sở giáo dục:* Những bình luận công khai liên quan đến trải nghiệm học tập của học viên nhắm đến các cơ sở giáo dục trên các nền tảng trực tuyến cũng là một dạng phản hồi. Việc ứng dụng mô hình kết quả nghiên cứu có thể giúp dễ dàng đánh giá và so sánh chất lượng của các cơ sở giáo dục khác, dựa trên dữ liệu phản hồi thực tế từ người học. Kết hợp đồng thời phương pháp Nhận diện thực thể đặt tên, các đánh giá sẽ được xác định tự động cụ thể tên của cơ sở đề cập trong bình luận. Các thông tin liên quan về chủ đề và cực cảm xúc sẽ được gắn trực tiếp với từng cơ sở, cho phép quá trình đánh giá trở nên rõ ràng và chính xác hơn. Nhờ vậy, các cơ sở giáo dục sẽ được so sánh với nhau theo cùng các các khía cạnh cụ thể để rút ra thế mạnh và hạn chế của từng bên. Kết quả phân tích sẽ là nền tảng cho các kế hoạch cải tiến, học hỏi từ các cơ sở khác, từ đó thúc đẩy sự cạnh tranh lành mạnh trong việc nâng cao chất lượng giáo dục và tạo động lực nâng cao hình ảnh trên thị trường giáo dục.

- *Quản lý danh tiếng cơ sở giáo dục:* Những bình luận trực tuyến có tác động đáng kể đối với danh tiếng của một cơ sở giáo dục. Thông qua việc tự động phân loại của mô hình, nhà quản lý có thể nhanh chóng đánh giá tình hình dư luận về cơ sở giáo dục trên các nền tảng trực tuyến, biết được khía cạnh nào đang được cộng đồng đánh giá cao và khía cạnh nào đang là nguyên nhân gây bất mãn. Từ đó, các hành động phù hợp được đưa ra nhằm duy trì và nâng cao hình ảnh của cơ sở. Khi xu hướng tiêu cực tăng cao, các biện pháp xử lý có thể kịp thời đưa ra, tránh sự lan rộng và gây tổn hại đến danh tiếng của cơ sở. Ngược lại, hoạt động quảng bá những điểm mạnh sẽ được thúc đẩy khi phát hiện xu hướng tích cực từ học viên nhằm thu hút thêm sự quan tâm từ cộng đồng cũng như các học viên tiềm năng. Bằng cách liên tục phân tích các bình luận mới, mô hình sẽ hỗ trợ xác định những xu hướng dài hạn, giúp nhà quản lý điều chỉnh chiến lược quản lý danh tiếng của cơ sở một cách linh hoạt và hiệu quả hơn.

- *Hỗ trợ người học hiệu quả:* Các phản hồi được phân loại bởi mô hình sẽ giúp hệ thống hiểu rõ sở thích từng học viên, cung cấp thông tin chi tiết về những điểm mạnh và điểm yếu mà mỗi học viên đã trải qua trong quá trình học tập. Đây sẽ là cơ sở để hệ thống cá nhân hóa nội dung đến từng người học một cách hiệu quả. Theo đó, phân tích các đánh giá khóa học của mỗi học viên sẽ giúp đo lường mức độ hài lòng người học theo chi tiết từng khía cạnh. Nhờ vậy, các khóa học tương tự hoặc được đánh giá tốt hơn bởi người học khác có thể sẽ được đề xuất. Từ đó, học viên có thể tiếp cận với những nội dung học tập phù hợp hơn và đảm bảo rằng học viên có trải nghiệm tích cực và hài lòng hơn trong quá trình học. Ngoài ra, các bài viết hoặc giải đáp phù hợp với phạm vi mà học viên

đang quan tâm cũng là một phần đề xuất mà hệ thống có thể gửi để hỗ trợ người học. Điều này không chỉ giúp học viên vượt qua những khó khăn trong học tập mà còn nâng cao khả năng tự học và tự nghiên cứu.

5. Kết luận

Nghiên cứu này đã trình bày một kiến trúc mô hình mới về SA phản hồi tiếng Việt của học viên. Kết quả thực nghiệm trên bộ dữ liệu UIT-VSFC cho thấy, sự hiệu quả của kiến trúc đa kênh đa mạng so với đơn kênh đơn mạng và tuần tự đa mạng phổ biến. Ngoài ra, kết quả thực nghiệm còn được so sánh với nghiên cứu gần nhất, chứng minh sự cải thiện đáng kể trong việc giải quyết vấn đề mất cân bằng. Hai mô hình phân loại tại hai nhiệm vụ đã được xây dựng, có ý nghĩa ứng dụng to lớn đối với các cơ sở giáo dục trong việc cải tiến chất lượng, quản lý danh tiếng, điểm chuẩn, hỗ trợ người học,… Có thể thấy, SA được ứng dụng vào những phản hồi học viên theo hướng tiếp cận học sâu là giải pháp cần thiết nhằm hỗ trợ nâng cao mức độ hài lòng người học. Đồng thời, việc tận dụng các bộ dữ liệu sẵn có để đào tạo mô hình cũng giúp các cơ sở cắt giảm chi phí thu thập và gán nhãn bộ dữ liệu mới, mang lại hiệu quả nguồn lực. Nhìn chung, nghiên cứu này đã giải quyết các hạn chế được đặt ra.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế tồn tại, tạo ra những khoảng trống trong hoạt động nghiên cứu. Về hiệu suất, chỉ số F1-Score (Macro) mặc dù được nâng cao đáng kể nhưng chỉ số F1-Score (Weighted) lại bị suy giảm nhẹ. Đây có thể được xem là một đánh đổi chấp nhận nhưng cũng có thể xem là một điểm hạn chế của mô hình nghiên cứu. Về nhiệm vụ, mô hình hiện tại chỉ có khả năng phân loại một nhãn cho một bình luận tại mỗi nhiệm vụ. Vấn đề này khiến cho mô hình nghiên cứu chưa tối ưu đối với các phản hồi đề cập nhiều chủ đề cùng lúc. Mô hình sẽ đòi hỏi các phản hồi cần được phân tách thành từng câu riêng lẻ trước khi phân tích nhưng điều này vẫn chưa thể giải quyết hoàn toàn các trường hợp một câu đề cập nhiều chủ đề. Về kiến trúc mô hình, nghiên cứu xây dựng hai mô hình riêng lẻ cho mỗi nhiệm vụ khác nhau, yêu cầu hai quá trình đào tạo riêng biệt. Hướng tiếp cận đa mô hình mặc dù giúp đạt được các tham số phù hợp nhất ở từng nhiệm vụ nhưng làm gia tăng đáng kể nổ lực tính toán trong quá trình đào tạo, đây cũng là một hạn chế. Về ứng dụng, kết quả nghiên cứu chỉ dừng lại ở mô hình và các hướng ứng dụng chỉ ở đề xuất. Do đó, nghiên cứu vẫn chưa triển khai sâu vào trong các cơ sở giáo dục, thiếu giá trị thực tế cho cơ sở giáo dục.

Trong tương lai, các nghiên cứu sẽ tập trung giải quyết các hạn chế còn tồn tại. Hiệu suất của mô hình không chỉ tại F1-Score (Macro) mà cả F1-Score (Weighted) sẽ được chú trọng đồng thời, các kiến trúc mô hình mới sẽ được nghiên cứu để cải thiện hiệu suất. Đặc biệt là F1-Score (Macro) vẫn còn nhiều tiềm năng cải thiện khi số lượng nghiên cứu liên quan còn rất hạn chế. Ngoài ra, mô hình SA dựa trên khía cạnh sẽ là một chủ đề đáng quan tâm cho các nghiên cứu tương lai khi có khả năng phân tích hiệu quả một câu đề cập nhiều chủ đề hay khía cạnh. Kết hợp với hướng tiếp cận đa nhiệm (multi-task), hai nhiệm vụ riêng sẽ có thể dễ dàng kết hợp trong một mô hình duy nhất, giúp giảm tham số, giảm đào tạo và giảm tài nguồn lực. Đồng thời, các nghiên cứu triển khai ứng dụng mô hình cũng sẽ được thực hiện để tiếp bước các mô hình kết quả đạt được. Các ứng dụng sẽ hướng đến các đề xuất đã nêu với khả năng xử lý dữ liệu lớn và xử lý thời gian thực, phù hợp với đặc điểm và sự phát triển của công nghệ hiện nay. Ngoài việc giải quyết các hạn chế, các nghiên cứu tương lai sẽ chú trọng thực nghiệm trên nhiều bộ dữ liệu khác nhau, nhằm gia tăng giá trị cho nghiên cứu khoa học. Bên cạnh đó, các dữ liệu phản hồi mới từ 2018 trở về sau sẽ được bổ sung vào bộ dữ liệu UIT-VSFC, giúp cập nhật những đặc điểm và xu hướng mới, tăng tính ứng dụng của bộ dữ liệu này. Đồng thời, lĩnh vực nghiên cứu sẽ chuyên sâu hơn, với các mô hình và dữ liệu dành riêng cho trung tâm giáo dục, trường trung học, đại học/cao đẳng, nền tảng giáo dục trực tuyến,… tăng tính phù hợp với từng nhóm.

Tài liệu tham khảo

Alaparthi, S., & Mishra, M. (2021). BERT: a sentiment analysis odyssey. *Journal of Marketing Analytics*, *9*, 118-126. https://doi.org/10.1057/s41270-021-00109-8

Cach, D. N. , Moreno-García, M. N., De la Prieta, F., Kien, N. V., & Vuong, N. M. (2023). Sentiment analysis for vietnamese - based hybrid deep learning models. In P. G. Bringas et al. (Eds.), *Proceedings of Hybrid Artificial Intelligent Systemspp* (Vol. 14001, pp. 293-303). Salamanca, Spain. https://doi.org/10.1007/978-3-031-40725-3_25

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches, *Proceedings of SSST-8 Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103-111). Doha, Qatar. https://doi.org/10.3115/v1/W14-4012

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster (2nd ed.). Simon and Schuster. https://books.google.com.vn/books?id=mjVKEAAAQBAJ

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* arXiv. https://doi.org/10.48550/arXiv.1412.3555

Das, J. K., Das, A., & Rosak-Szyrocka, J. (2022). A Hybrid Deep Learning Technique for Sentiment Analysis in E-Learning Platform with Natural Language Processing, *Proceedings of 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)* (pp. 1-7). Split, Croatia. https://doi.org/10.23919/SoftCOM55329.2022.9911232

Dat, N. Q., & Anh, N. T. (2020). *PhoBERT: Pre-trained language models for Vietnamese*. arXiv. https://doi.org/10.48550/arXiv.2003.00744

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Dyulicheva, Y. Y., & Bilashova, E. A. (2021). Learning Analytics of MOOCs based on natural language processing, *Proceedings of 4th Workshop for Young Scientists in Computer Science & Software* (pp. 187-197). Kryvyi Rih, Ukraine. https://ceur-ws.org/Vol-3077/paper15.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press. https://books.google.com.vn/books?id=omivDQAAQBAJ

Nguyen, H. Q., Vu, L., & Nguyen, Q. U. (2020). Residual Attention Bi-directional Long Short-term Memory for Vietnamese Sentiment Classification. *Journal of Science and Technique-Section on Information and Communication Technology*, *9*(02). https://doi.org/10.56651/lqdtu.jst.v9.n02.212.ict

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huy, H. D., Hang, D. T. T., Kiet, N. V., & Ngan, N. T. L., (2020). A simple and efficient ensemble classifier combining multiple neural network models on social media datasets in Vietnamese. In N. L. Minh, L. C. Mai, & Song, S. (Eds.). *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, PACLIC 2020* (pp. 420-429), Hanoi, Vietnam. Association for Computational Linguistics. https://aclanthology.org/2020.paclic-1.48/

Kandhro, I. A., Wasi, S., Kumar, K., Rind, M., & Ameen, M. (2019). Sentiment Analysis of Students' Comment by using Long-Short Term Model. *Indian Journal of Science and Technology*, *12*(8), 1-16. https://doi.org/10.17485/ijst/2019/v12i8/141741

Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A Systematic mapping study. *Applied Sciences*, *11*(9). https://doi.org/10.3390/app11093986

Kastrati, Z., Imran, A. S., & Kurti, A. (2020). Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs. *IEEE Access*, *8*, 106799-106810. https://doi.org/10.1109/ACCESS.2020.3000739

Kiet, N. V., Vu, N. D., Phu, N. V. X., Tham, T. H. T., & Ngan, N. T. L. (2018). UIT-VSFC: Vietnamese Students' feedback corpus for sentiment analysis, *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 19-24). Ho Chi Minh City, Vietnam. https://doi.org/10.1109/KSE.2018.8573337

Lac, L. S., Thin, D. V., Ngan, N. T. L., & Son, T. Q. (2020). *A multi-filter BiLSTM-CNN architecture for Vietnamese sentiment analysis*. In M. Hernes, K. Wojtkiewicz, & E. Szczerbicki (Eds.), Advances in Computational Collective Intelligence. ICCCI 2020. Communications in Computer and Information Science (vol 1287, pp. 752–763). Springer, Cham. https://doi.org/10.1007/978-3-030-63119-2_61

Liu, B. (2022). *Sentiment Analysis and Opinion Mining*. Springer Nature. https://doi.org/10.1007/978-3-031-02145-9

Loc, C. V., Viet, T. X., Viet, T. H., Thao, L. H., & Viet, N. H. (2022). A Text Classification for Vietnamese Feedback via PhoBERT-Based Deep Learning. In X. S. Yang, S. Sherratt, N. Dey, A. Joshi (Eds.), *Proceedings of Seventh International Congress on Information and Communication Technology* (pp 259-272). Springer, Singapore. https://doi.org/10.1007/978-981-19-2394-4_24

Loc, T. T., Linh, L. H., & Phuc, L.D. T. (2020). Sentiment analysis implementing BERT-based pre-trained language model for Vietnamese. *Proceedings of 2020 7th NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 362-367). Ho Chi Minh City, Vietnam. https://doi.org/10.1109/NICS51282.2020.9335912

Phu, N. V. X., Tham, H. T. T., Kiet, N. V., & Ngan, N. T. L. (2019). Deep learning versus traditional classifiers on Vietnamese students' feedback corpus. *Proceedings of 2018 5th NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 75-80), Ho Chi Minh City, Vietnam. https://doi.org/10.1109/NICS.2018.8606837

Vu, N. D., Kiet, N. V., & Ngan, N. T. L. Vu(2018). Variants of long short-term memory for sentiment analysis on Vietnamese students' feedback corpus. *Proceedings of 10th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 306-311). Ho Chi Minh City, Vietnam. https://doi.org/10.1109/KSE.2018.8573351

Onan, A. (2020). Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education*, *28*(1), 117-138. https://doi.org/10.1002/cae.22179

Peng, H., Zhang, Z., & Liu, H. (2022). A sentiment analysis method for teaching evaluation texts using attention mechanism combined with CNN-BLSTM model. *Scientific Programming*. https://doi.org/10.1155/2022/8496151

Quan, V. H., Huy, N. T., Bac, L., & Minh, N. L. (2017). Multi-channel LSTM-CNN model for Vietnamese sentiment analysis. *Proceedings of 2017 9th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 24-29). Hue, Vietnam. https://doi.org/10.1109/KSE.2017.8119429

Quynh, D. V. X., Laosen, K., & Laosen, N. (2021). An evaluation of the UIT-VSFC Dataset using modern machine learning techniques and word embeddings, *Proceedings of 2021 25th International Computer Science and Engineering Conference (ICSEC)* (pp. 394-399). Chiang Rai, Thailand. https://doi.org/10.1109/ICSEC53205.2021.9684597

Razinkina, E., Pankova, L., Trostinskaya, I., Pozdeeva, E., Evseeva, L., & Tanova, A. (2018). Student satisfaction as an element of education quality monitoring in innovative higher education institution. *E3S Web of Conferences*, *33*. https://doi.org/10.1051/e3sconf/20183303043

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *2*. https://doi.org/10.1016/j.nlp.2022.100003

Sindhu, I., Daudpota, S. M., Badar, K., Bakhtyar, M., Baber, J., & Nurunnabi, M. (2019). Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation. *IEEE Access*, *7*, 108729-108741. https://doi.org/10.1109/ACCESS.2019.2928872

Sutoyo, E., Almaarif, A., & Yanto, I. T. R. (2021). Sentiment analysis of student evaluations of teaching using deep learning approach. In J. H. Abawajy, K. K. R. Choo, & H. Chiroma (Eds.), *Proceedings of International Conference on Emerging Applications and Technologies for Industry 4.0 (EATI'2020)*. Springer, Cham. https://doi.org/10.1007/978-3-030-80216-5_20

Thang, P. C., Nam, N. Q., Thanh, D. C., Hop, D. T., & Kiet, N. V. (2023). ViCGCN: Graph convolutional network with contextualized language models for social media mining in Vietnamese. *arXiv*. https://doi.org/10.48550/arXiv.2309.02902

Thanh, V., Dat, N. Q., Dai, N. Q., Dras, M., & Johnson, M. (2018). VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In Y. Liu, T. Paek, M. Patwardhan (Eds.). *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 56-60). New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1801.01331

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780. https://doi.org/10.1007/s10462-022-10144-1

Zheng, J., Wang, J., Ren, Y., & Yang, Z. (2020). Chinese sentiment analysis of online education and internet buzzwords based on BERT. *Journal of Physics: Conference Series*, *1631*. https://doi.org/10.1088/1742-6596/1631/1/012034