Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), TP. HCM, ngày 23-24/12/2021
DOI: 10.15625/vap.2021.0088

PHÂN TÍCH QUAN ĐIỂM XÃ HỘI ĐỐI VỚI ĐẠI HỌC PHAN THIẾT

Võ Quốc Tuấn[1], Trần Thanh Phước[2], Trần Thanh Trâm[3]

[1]Phòng Quản lý Đào tạo, Trường Đại học Phan Thiết, Tp. Phan Thiết, Bình Thuận
[2] Phòng Lab Natural Language Processing and Knowledge Discovery, Khoa Công nghệ thông tin, Trường Đại học Tôn Đức Thắng, Thành phố Hồ Chí Minh
[3]Phòng Quản lý khoa học và Đào tạo Sau đại học, Trường Đại học Công nghiệp thực Phẩm, Thành phố Hồ Chí Minh

vqtuan@upt.edu.vn, tranthanhphuoc@tdtu.edu.vn, tramtt@hufi.edu.vn

TÓM TẮT: *Trong giáo dục hiện đại, trường đại học đóng vai trò là nơi cung cấp dịch vụ, học sinh sinh viên và phụ huynh là khách hàng. Việc nắm bắt được cảm xúc, quan điểm của những đối tượng khách hàng này (gọi chung là xã hội) đối với các dịch vụ mà trường học cung cấp là vô cùng cần thiết. Qua đó, trường học sẽ tiếp nhận những mặt tích cực lẫn tiêu cực để từ đó phát huy mặt tích cực và hạn chế mặt tiêu cực. Trong phạm vi bài báo này, chúng tôi tập trung vào hai việc: (1) Xây dựng bộ dữ liệu cảm xúc của xã hội đối với Trường Đại học Phan Thiết bao gồm 3 nhãn: tích cực, trung tính và tiêu cực; (2) Đề xuất sử dụng các mô hình học sâu như CNN, LSTM, BERT và PhoBERT để thử nghiệm cho bài toán phân tích quan điểm của Đại học Phan Thiết. Kết quả thử nghiệm cho thấy, PhoBERT cho kết quả cao nhất trên bộ dữ liệu Đại học Phan Thiết với F1-score là 89,68%.*

Từ khóa: *Phân tích quan điểm, khai phá dữ liệu giáo dục, phân loại văn bản, pretrain BERT, PhoBERT.*

I. GIỚI THIỆU

Trường Đại học Phan Thiết (UPT: University of Phan Thiet) được thành lập tháng 03/2009 theo Quyết định số 394/2009/QĐ-TTg, ngày 25 tháng 3 năm 2009 của Thủ tướng Chính phủ. Trong quá trình hoạt động đào tạo, hằng năm nhà trường khảo sát, thu thập các ý kiến, đánh giá về chất lượng đào tạo, cơ sở vật chất, chăm sóc sinh viên,… Các hình thức khảo sát ý kiến sinh viên vào cuối học kỳ trên hệ thống xem kết quả môn học, khảo sát các ý kiến cựu sinh viên, học viên cao học, doanh nghiệp, trang fanpage UPT, fanpage của các khoa, UPT confessions. Việc thu thập thông tin này đã giúp cho nhà trường rất nhiều trong quá trình hoàn thiện hơn về công tác đào tạo, các hoạt động cộng đồng và cơ sở vật chất nhằm phục vụ cho việc giảng dạy được tốt hơn. Đánh giá và phân loại những ý kiến từ các khảo sát trên đã mang lại kết quả tốt hơn cho người học lẫn doanh nghiệp và toàn xã hội.

Như vậy, vấn đề làm thế nào để có thể khai thác được những thông tin của tất cả các ý kiến của sinh viên, cựu sinh viên, học viên cao học, doanh nghiệp,… đánh giá và phân loại những ý kiến theo từng ngành học trở nên khả thi và mang lại kết quả tốt hơn cho người học lẫn doanh nghiệp và toàn xã hội.

Phân tích quan điểm là một dạng của bài toán phân loại văn bản dựa trên văn bản ngôn ngữ tự nhiên nhằm phát hiện ra thái độ, màu sắc tình cảm của người dùng thông qua bình luận (comment) trên các trang phim, ca nhạc, facebook, twitter, các kênh khảo sát trực tuyến nhằm đánh giá về một sản phẩm, hoạt động đào tạo đối với một trường đại học, ví dụ:

- Trường có view nhìn ra biển rất đẹp (tích cực).
- Phòng 102 âm thanh bị rè không nghe rõ (tiêu cực).
- Chúng tôi không có ý kiến (trung tính).

Đầu vào của bài toán Phân tích quan điểm là một câu hay một đoạn văn bản ngắn, đầu ra là các giá trị xác suất của nhiều lớp quan điểm mà ta cần xác định. Trong nghiên cứu này tôi chọn loại bài toán phân tích quan điểm thành 3 lớp: tích cực (positive), tiêu cực (negative) và trung tính (neutral). Chúng tôi tập trung chủ yếu vào bài toán phân tích quan điểm trên phạm vi dữ liệu quan trọng đó là giáo dục.

Để giải quyết bài toán này, điều đầu tiên chúng ta cần phải có chính là kho ngữ liệu phục vụ cho thực nghiệm. Hiện tại kho ngữ liệu phân tích quan điểm cho UPT là chưa có. Vì vậy, việc đầu tiên cần thực hiện cho công trình này là thu thập kho ngữ liệu đủ lớn cho việc thực nghiệm. Chúng tôi sẽ trình bày chi tiết các bước thu thập, tiền xử lý dữ liệu,… ở Phần IV.B. Sau đó, chúng tôi sử dụng nhiều mô hình tiên tiến để thực nghiệm trên bộ dữ liệu vừa thu thập được, bao gồm các mô hình như: LSTM, CNN, các mô hình BERT, PhoBERT. Kết quả thực nghiệm cho thấy PhoBERT là mô hình cho kết quả tốt nhất trên kho ngữ liệu giáo dục của chúng tôi. Trong khi đó, mô hình LSTM cho kết quả thấp nhất.

Phần còn lại của bài báo được trình bày theo cấu trúc sau: Phần II và phần III lần lượt trình bày các công trình liên quan cũng như một số kiến thức nền tảng của bài báo. Các bước thu thập dữ liệu cũng như mô hình chi tiết của bài toán phân tích quan điểm của UPT được trình bày ở Phần IV. Trong khi đó, phần thực nghiệm và kết luận lần lượt được trình bày ở Phần V và VI.

II. CÔNG TRÌNH LIÊN QUAN

J. T. Mendez và cộng sự [1] đã nghiên cứu về sự hài lòng người sử dụng giao thông công cộng tại thành phố Santiago, Chile thông qua mạng xã hội twitter. Các kỹ thuật khai thác văn bản, phân tích quan điểm và mô hình hóa chủ đề, sử dụng các câu hỏi trong nghiên cứu đánh giá mức độ hài lòng của người sử dụng giao thông công cộng đặc biệt là xe

buýt. Một nghiên cứu về việc kết hợp các phương pháp phân tích quan điểm dựa trên bộ từ điển và học máy cho các đánh giá của khách hàng trong tiếng Việt được Son Trinh [2] và cộng sự triển khai nghiên cứu. Tác giả sử dụng những dấu hiệu cảm xúc và giá trị của cảm xúc là những thông tin được trích xuất từ tập dữ liệu gốc. Chúng còn được gọi là các tính năng, được sử dụng để phân loại cảm xúc. Để huấn luyện (train) cho quá trình phân loại, các tập dữ liệu huấn luyện cần được chuyển thành một véctơ chứa các tính năng đó, gọi là véctơ đặc trưng. Việc phân tích quan điểm văn bản trong tiếng Việt đã có một số nghiên cứu. Đặc biệt, H. Nam Nguyen và cộng sự [3] đưa ra vấn đề xây dựng từ điển cảm xúc trong tiếng Việt là rất khó và mất thời gian. Cách tiếp cận khai phá ý kiến cộng đồng bằng ngôn ngữ tiếng Việt bằng cách sử dụng từ điển cảm xúc trong các lĩnh vực cụ thể để cải thiện độ chính xác. Tran Sy Bang [4] và cộng sự đề xuất một kỹ thuật phân tích quan điểm cho các văn bản tiếng Việt dựa trên thuật ngữ tiếp cận lựa chọn tính năng đặc trưng, dùng 3 thuật toán Naive Bayes (NB), cấy quyết định và máy véctơ hỗ trợ (SVM) để phân loại văn bản.

Từ khi BERT [5] ra đời đánh dấu vượt bật trong mảng xử lý ngôn ngữ tự nhiên, hàng loạt các nghiên cứu dựa trên BERT cho bài toán phân tích quan điểm ra đời cải thiện đáng kể, tinh chỉnh BERT cho bài toán phân loại văn bản được nhóm Quoc Thai Nguyen và cộng sự [6] đã đưa ra các mô hình kết hợp BERT với CNN, LSTM, RCNN trên bộ dữ liệu tiếng Việt được phân loại thành hai nhãn tích cực và tiêu cực đạt kết quả F1-score 91,15%.

Gần đây, nghiên cứu phân tích ý kiến tiếng Việt dành cho mảng giáo dục được chú trọng, thu thập ý kiến khảo sát và phản hồi thông tin trong giáo dục để cải tiến chất lượng giảng dạy và quy trình quản lý [7], quan tâm đến ý kiến sinh viên được thu thập trong các cuộc khảo sát, giới thiệu khái niệm mới gọi là Bag-of-Structure (BOS), các thuật toán học máy để khai thác cảm xúc và phân loại. Kiet Van Nguyen và cộng sự [8] đã xây dựng bộ dữ liệu hơn 16.000 câu ý kiến phản hồi của sinh viên của trường đại học đã được gán nhãn tích cực, tiêu cực hay trung tính với độ chính xác lên đến 87,94%. Bên cạnh đó, Phu X.V. Nguyen và cộng sự [9] đã dựa vào bộ dữ liệu trên và sử dụng các thuật toán Deep Learning NB, LSTM, Bi-LSTM, so sánh các thuật toán nêu trên Bi-LSTM đạt kết quả cao nhất với F1-score 89,6%.

Ý kiến đánh giá quá trình đào tạo thường bình luận về nhiều khía cạnh khác nhau, ví dụ chất lượng của một môn học và khả năng ứng dụng chúng vào thực tế, hay là chất lượng về phục vụ cộng đồng đối với một trường đại học. Phản hồi của sinh viên hỗ trợ cải tiến trong giảng dạy nhấn mạnh các vấn đề khác nhau mà các sinh viên có được bài giảng từ giảng viên [8], [10]. Ví dụ về vấn đề này khi một sinh viên không hiểu một phần của bài giảng hoặc một ví dụ cụ thể. Phản hồi của sinh viên có thể làm sáng tỏ những gì sinh viên làm, không hiểu, những gì họ thích và không thích về bài giảng, ví dụ: tốc độ giảng dạy của giảng viên quá nhanh hoặc quá chậm. Ngoài ra, thông tin phản hồi của sinh viên có thể giúp giảng viên hiểu được hành vi học tập của sinh viên. Xây dựng hệ thống để khai thác phản hồi của sinh viên cũng là cách tích cực để cải thiện việc học tập của sinh viên [11]. Sử dụng các ý kiến quan điểm người dùng trên mạng xã hội Twitter để đánh giá các trường đại học [12] cũng là một trong những phương pháp hiệu quả. Phản hồi thường được thu thập vào cuối học kỳ hoặc khóa học, sinh viên được yêu cầu làm nổi bật các vấn đề khác nhau có thể xảy ra trong khóa học [13]. Tuy nhiên, phản hồi của sinh viên có lợi hơn khi được thực hiện trong thời gian thực, vì nó cho phép các giảng viên giải quyết kịp thời các vấn đề.

Hiện tại ở Việt Nam, phân tích quan điểm về giáo dục rất ít, đặc biệt về lĩnh vực giáo dục đại học. H T. Vo và cộng sự [14] nghiên cứu phân loại chủ đề và phân tích quan điểm cho hệ thống khảo sát giáo dục Việt Nam. Những ý kiến phản hồi từ nơi mà các sinh viên thực tập, các thông tin thu thập như: lớp học, chất lượng luận án,… được phân tích giới hạn ở mức câu được phân lớp tích cực, tiêu cực. Chi tiết hơn Kiet Van Nguyen và cộng sự [8] đã nghiên cứu những ý kiến của sinh viên khi xem kết quả môn học của một trường đại học và được phân lớp tích cực, tiêu cực hay trung tính.

III. CƠ SỞ LÝ THUYẾT

A. Convolutional Neural Network (CNN)

Mạng CNN là một tập hợp các lớp tích chập (convolution) chồng lên nhau sử dụng các hàm phi tuyến tính như ReLU hay Tanh để kích hoạt trọng số tại mỗi node. Mỗi lớp thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng các cho các lớp tiếp theo. CNN không những được áp dụng thành công trong lĩnh vực xử lý ảnh, mà còn được sử dụng thành công trong một số bài toán phân loại văn bản. Trong [15], tác giả cho thấy rằng sử dụng CNN đơn giản điều chỉnh siêu tham số và sử dụng pre-train word2vec cho kết quả cực tốt, giải pháp cải thiện kết quả trong đó có phân loại cảm xúc và phân loại văn bản.

B. Long Short Term Memory (LSTM)

Long Short Term Memory là một mạng cải tiến của mạng RNN truyền thống, nhằm giải quyết vấn đề học xa. LSTM được giới thiệu bởi Hochreiter & Schmidhuber 1997 [16] được sử dụng nhiều nhất cho phân loại văn bản. Tất cả các mạng hồi quy đều chỉ là các chuỗi các môđun lặp đi lặp lại. Trong mạng RNN (Recurrent Neural Networ) chuẩn, các môđun này có kiến trúc đơn giản với một tầng tanh. Thay vì chỉ có một tầng đơn, chúng có 4 tầng ẩn (3 sigmoid và 1 tanh) tương tác với nhau.

C. BERT (Bidirectional Encoder Representations from Tranformers)

1. Ý nghĩa của BERT

BERT [5] được hiểu là pre-train model (mô hình học sẵn) được các nhà nghiên cứu tại Google AI Language phát triển. BERT được thiết kế đào tạo ra các véctơ đại diện cho ngôn ngữ văn bản thông qua ngữ cảnh 2 chiều trái và

phải trong tất cả các lớp. Kết quả là, véctơ được sinh ra từ mô hình BERT được tinh chỉnh (Fine-tuning) với các lớp đầu ra bổ sung đã tạo ra nhiều kiến trúc đáng kể trong nhiệm vụ xử lý ngôn ngữ tự nhiên: hỏi đáp, suy luận ngôn ngữ.

2. Các thành phần chính của BERT

Kiến trúc Transformers được nhắc đến trong nghiên cứu của Vaswani và cộng sự [17] gồm 2 thành phần chính Encoder và Decoder.

Encoder: Gồm các module giống nhau được xếp chồng lên nhau, mỗi module này được chia thành 2 lớp con: (1) là cơ chế tự chú ý nhiều đầu (multi-head self-attention), (2) một mạng truyền thẳng kết nối đầy đủ (fully connected feed-forward network). Số module được xếp chồng lên nhau có thể tùy biến.

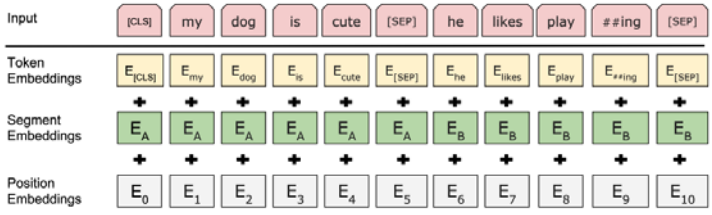
Decoder: Có kiến trúc giống hệt Encoder, tuy nhiên ở giữa hai module con, một lớp tự chú ý (self-attention) nhiều đầu được gắn thêm vào với mục đích giúp Decoder sử dụng đầu ra của Encoder với attention. Ngoài ra, lớp self-attention được chỉnh sửa lại để đảm bảo các từ dự đoán cho vị trí i chỉ có thể phụ thuộc vào các đầu ra tại các vị trí nhỏ hơn i. Điều này bảo đảm rằng việc dự đoán một từ dựa vào các từ trước đó.

Tác vụ huấn luyện (pretraining task): Sự thành công của BERT sử dụng hai nhiệm vụ dự đoán không giám sát là: Masked Language Model (MLM) và dự đoán câu tiếp theo (Next Sentence Prediction).

MLM (Masked Language Model): BERT che giấu (mask out) một số token trong chuỗi một cách ngẫu nhiên, sau đó dùng các từ còn lại để dự đoán những từ bị che giấu, quá trình này gọi MLM.

Dự đoán câu tiếp theo (Next Sentence Prediction): Một số nhiệu vụ như trả lời tự động (Question Answering) trong NLP yêu cầu hiểu biết dựa vào mối liên hệ giữa hai câu với nhau. Vì vậy, cần xây dựng một mô hình dự đoán câu tiếp theo của câu trước đó hay không.

Hình *1* thể hiện một ví dụ về biểu diễn đầu vào của BERT ứng với câu tiếng Anh "my dog is cute. He likes playing."

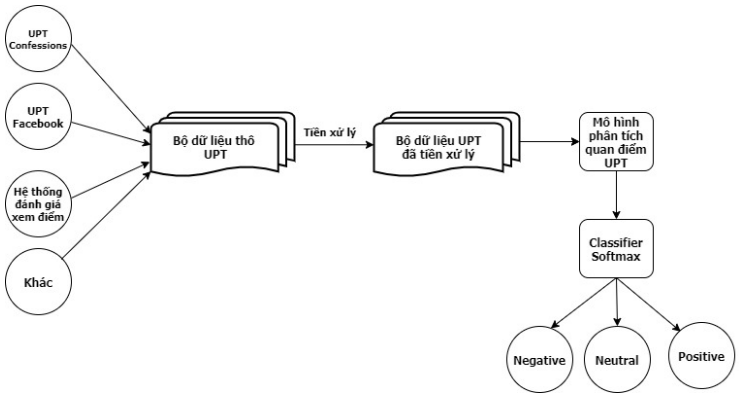


Hình 1. Biểu diễn đầu vào của BERT [5]

IV. MÔ HÌNH PHÂN TÍCH QUAN ĐIỂM XÃ HỘI ĐỐI VỚI UPT

A. Mô hình tổng quát

Hình 2 trình bày mô hình tổng quát cho bài toán phân tích quan điểm xã hội đối với UPT.



Hình 2. Mô hình tổng quát cho bài toán phân tích quan điểm xã hội đối với UPT

B. Xây dựng bộ dữ liệu UPT

1. Chúng tôi thực hiện thu thập dữ liệu từ nhiều nguồn khác nhau được liệt kê dưới đây:

- Fanpage UPT, fanpage các khoa, UPT Confessions.
- Ý kiến đánh giá, nhận xét trên hệ thống xem điểm trực tuyến của sinh viên từ năm 2016 - 2019, học viên cao học.
- Các ý kiến nhận xét của doanh nghiệp trong ngày hội việc làm tổ chức hàng năm hoặc doanh nghiệp nơi sinh viên đang công tác.

- Các ý kiến phụ huynh của sinh viên, cựu sinh viên, khảo sát qua Google docs do Phòng Khảo thí - Đảm bảo chất lượng và Thanh tra.

2. Tiền xử lý dữ liệu:

Chúng tôi thực hiện các bước tiền xử lý thật kỹ để tăng hiệu suất cho mô hình được chính xác nhất. Dữ liệu thu thập từ các fangpage liên quan đến UPT ất nhiều từ ngữ lóng, các kí hiệu, các từ viết tắt,… Vì vậy, chúng tôi tiến hành việc xử lý được mô tả dưới đây:

- Loại bỏ các HTML tag.
- Loại bỏ những ký tự lặp lại nhiều lần khi chúng không phải là alphanumeric. Ví dụ: "…" -> "???"->"?". Việc này cũng loại bỏ những khoảng trắng thừa và giúp cho bộ phận tách từ hoạt động hiệu quả hơn. Đối với câu văn dài thì chúng tôi giữ lại dấu "." để phân biệt giữa các câu trong câu văn dài.
- Trong những câu bình luận, việc sai lỗi chính tả, từ viết tắt rất nhiều, chúng tôi thực hiện chỉnh sửa những lỗi chính tả: "Nguyễn Diệp Trân Trân ra nhìn xe mà k pk ns j. Chỉ pk tức vs tức vì ns ra cũg k có ai đứg ra giải quyết cho mình". Được sửa lại: "Nguyễn Diệp Trân Trân ra nhìn xe mà không biết nói sao. Chỉ biết tức với tức vì nói ra cũng không có ai đứng ra giải quyết cho mình".
- Loại bỏ những các ký hiệu biểu hiện emoji, Thay thế các ký tự "_" thành ký tự khoảng trắng.
- Xóa những ký tự đánh dấu đầu dòng như: "-", , IV,…
- Chuyển tất cả chữ hoa về chữ thường,…

Đối với ngôn ngữ tiếng Việt thì việc tách từ trong tiền xử lý là không thể thiếu bởi vì từ vựng trong tiếng Việt có thể được tạo thành từ một từ ngữ hoặc nhiều từ ngữ khác nhau ví dụ như từ vựng "sinh viên" được tạo thành từ 2 từ "sinh" và "viên", nhưng khi kết hợp 2 từ này lại thì nó lại mang ý nghĩa khác. Để giải quyết vấn đề này, chúng tôi sử dụng thư viện vncorenlp[1] để tách từ trong dữ liệu đầu vào.

3. Gán nhãn cảm xúc

Quá trình gán nhãn dữ liệu: Tôi cùng hai đồng nghiệp tham gia gán nhãn liệu gồm Lê Trung Thành, Lương Quốc Vũ (Ban Thông tin - Truyền thông & quản trị mạng), Chúng tôi tìm hiểu quy tắc gán nhãn và chia nhau mỗi người gán 100 câu văn vài lần cho đến sự đồng thuận giữa ba người đạt khoảng 85% trở lên. Dựa vào kết quả trên chúng tôi chia nhau như sau: Tôi phụ trách: 3000 câu văn, hai đồng nghiệp còn lại mỗi người 1500 câu văn. Sau khi quá trình gán nhãn được thực hiện, chúng tôi đã tổng hợp và đánh giá chất lượng của bộ dữ liệu theo độ đo đồng thuận Cohen's Kappa K[2] theo công thức $K = \frac{P_0 - P_1}{1 - P_e}$, trong đó P0 là độ đồng thuận quan sát được 95,27%, Pe là độ đồng thuận kỳ vọng 46,19%, K là độ đồng thuận 91,2%.

Giả sử có một câu phản hồi của sinh viên: "Hệ thống wifi ở trường rất chậm", nhiệm vụ là xác định có phải quan điểm của câu văn đó mang tính cảm xúc tiêu cực, tích cực hay trung tính. Bảng 1 thể hiện một số câu ví dụ trong bộ dữ liệu UPT.

Bảng 1. Một số câu bình luận trong bộ dữ liệu gán nhãn cảm xúc

STT	Câu văn	Nhãn cảm xúc
1	Trường nổi_trội về các ngành_nghề dịch_vụ du_lịch	Tích cực
2	Chúng_tôi không có ý_kiến gì cả	Trung tính
3	Hệ_thống wifi ở trường rất chậm	Tiêu cực

Tích cực: Những câu văn thể hiện sự hài lòng khi khảo sát ở doanh nghiệp có sinh viên làm việc, sự hài lòng của người học, những lời ngợi khen, sự động viên tích cực nhất dành cho người học, các tổ chức đào tạo giáo dục về các nội dung như: giảng viên, nội dung môn học, cơ sở vật chất, chương trình đào tạo và một số khác. Ví dụ, câu văn "Trường nổi trội về các ngành nghề dịch vụ du lịch" sẽ được gán nhãn cảm xúc tích cực.

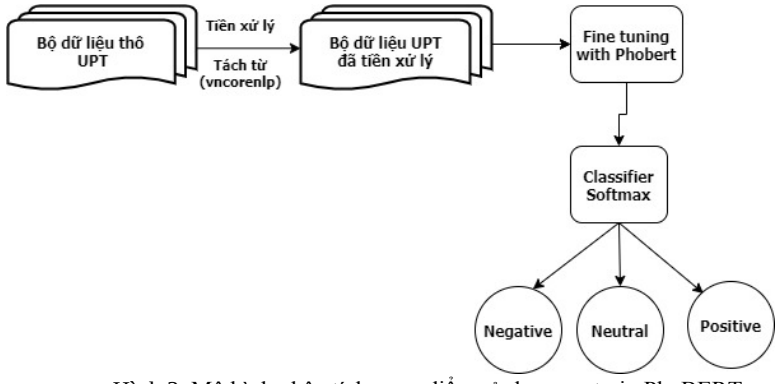
Trung tính: Những câu văn không hàm ý chứa bất kỳ cảm xúc nào, những câu văn không hoàn chỉnh, không rõ ràng về nghĩa, có ý nghĩa chung chung. Ví dụ, trong UPT Confession được ghi "Thực tế thì tùy vào ngành mà e chọn để định hướng nhé" được gán nhãn cảm xúc trung tính, câu văn này mang tính trao đổi nó như một cụm danh từ bình thường.

Tiêu cực: Những câu văn thể hiện sự không hài lòng của người sử dụng lao động, của người học, các tổ chức đào tạo. Những lời phàn nàn, không đồng ý, những lời đề nghị cần thiết nhất đối với nhà trường như giảng viên, môn học, chương trình đào tạo, cơ sở vật chất và một số khác. Ví dụ, câu văn "Hệ thống wifi trường rất chậm" sẽ được gán nhãn cảm xúc tiêu cực.

[1] https://github.com/vncorenlp/VnCoreNLP
[2] https://en.wikipedia.org/wiki/Cohen%27s_kappa

Trường hợp khiến ta khó khăn trong câu thể hiện vừa mang quan điểm tiêu cực và tích cực, điều này làm khó khăn khi gán nhãn. Chúng ta thường gặp một số câu có sử dụng từ liên kết như: nhưng, tuy nhiên, mặc dù, dù, dù rằng, tuy rằng và một số khác. Trong trường hợp này chúng tôi lựa chọn mệnh đề có tính phân cực mạnh hơn để gán nhãn. Ví dụ, câu văn "cần nâng cấp thư viện, phòng máy, phòng học, còn về môi trường xung quanh rất ồn" được gán nhãn tiêu cực, trong khi đó ta thấy "môi trường xung quanh rất ồn" mang cảm xúc tích cực.

C. Sử dụng pretrain PhoBERT cho bài toán phân tích quan điểm



Hình 3. Mô hình phân tích quan điểm sử dụng pretrain PhoBERT

Hình 3 có thể xem là bản cập nhật của Hình 2 thể hiện mô hình phân tích quan điểm đối với UPT sử dụng PhoBERT để huấn luyện. PhoBERT là một mô hình tiền huấn luyện cho tiếng Việt dựa trên kiến trúc RoBERTa do 2 tác giả D. Q. Nguyen và A. T. Nguyen [18] thuộc viện nghiên cứu VinAI Việt Nam được giới thiệu vào tháng 03/2020. PhoBERT cũng có 2 phiên bản PhoBERT_base với 12 transformers block và PhoBERT_large với 24 trainsformers block. PhoBERT được huấn luyện trên khoảng 20GB dữ liệu bao gồm 1GB dữ liệu Vietnamese Wikipedia corpus và 19 GB được thu thập và xử lý từ bộ dữ liệu thô 50 GB[3].

PhoBERT sử dụng RDRSegmenter của VncoreNLP để tách từ cho dữ liệu đầu vào trước khi đi qua BPE encoder. PhoBERT dựa trên kiến trúc RoBERTa bỏ đi nhiệm vụ dự đoán câu tiếp theo mà chỉ sự dụng mặt nạ.

Trong thử nghiệm của bài báo này, chúng tôi sử dụng PhoBERT_base để thực nghiệm trên bộ dữ liệu UPT. Mục tiêu của chúng tôi tạo ra một mô hình lấy một câu (giống như các câu trong tập dữ liệu), đầu ra cuối cùng của mô hình mang tính cảm xúc là tích cực, trung tính hay tiêu cực. Hình 4 là một ví dụ một câu văn được đưa vào tiền xử lý, tách từ, chuyển vào mô hình PhoBERT, hàm softmax dùng để phân loại kết quả mang tính cảm xúc.

Hình 4. Ví dụ cho bài toán phân tích quan điểm

Biểu diễn một câu văn vào mô hình phân tích quan điểm:

Chúng tôi thực hiện dự đoán một câu "khuôn viên Trường Đại học Phan Thiết đẹp thoáng mát" được mô tả Hình 5.

Bước 1: Sử dụng thư viện vncorenlp để thực hiện tách từ như sau: "khuôn_viên Trường Đại_học Phan_Thiết đẹp thoáng mát".

Bước 2: Thêm mã thông báo [CLS] để đánh dấu đầu câu và [SEP] để đánh dấu cuối câu.

Bước 3: Sử dụng thuật toán BPE (Byte Pair Encoding) để đưa câu đầu vào dưới dạng subword và ánh xạ các subword về dạng index trong từ điển.

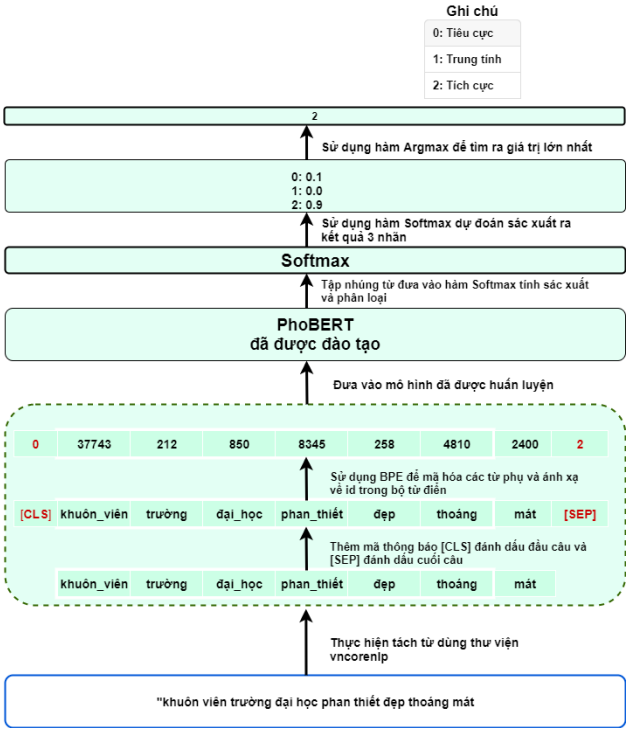
Bước 4: Chuyển vào mô hình PhoBERT Fine-tuning. Đầu ra là một véctơ đặc trưng, hàm softmax để tính xác suất đầu ra, hàm Argmax để chọn giá trị lớn nhất chọn ra giá trị cuối cùng.

Công thức của hàm Softmax:

Theo (1), trong đó z_i gồm các giá trị là phần tử của véctơ đầu vào X, nó có thể là con số âm hoặc dương, vậy $\exp(z_i)$ trả về kết quả trong phạm vi từ 0 đến 1. Tất cả các a_i dựa vào z_i và cho kết quả tổng là 1. Ta thấy trong Hình 5 hàm softmax trả về 3 lớp có giá trị: lớp 1: 0.1; lớp 0: 0.0; lớp 2: 0.9. Tiếp đến chúng tôi sử dụng hàm argmax để tìm giá trị lớn nhất cho kết quả là: 0.9 ở vị trí 2 là nhãn tích cực

[3] https://github.com/binhvq/news-corpus

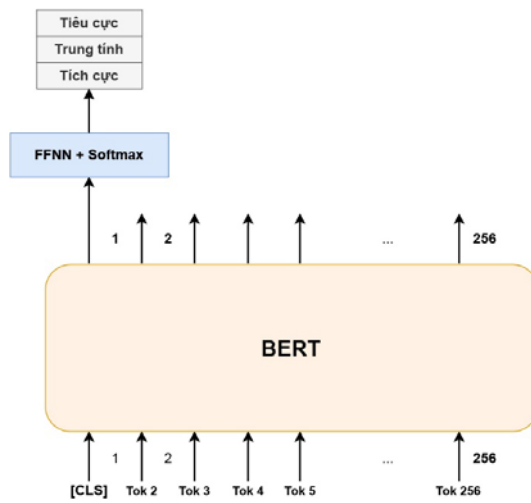
$$a_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}, \forall i = 1,2,\dots,C \quad (1)$$



Hình 5. Mô phỏng một câu bình luận để dự đoán nhãn cảm xúc

D. Mô hình tinh chỉnh BERT phân tích ý kiến cho tiếng Việt

Tinh chỉnh BERT- base cho bài toán phân tích ý kiến cho tiếng Việt chỉ sử dụng mã thông báo [CLS] ở vị trí đầu câu văn. Đầu ra véctơ của mã thông báo này chuyển đến một mạng truyền thẳng làm nhiệm vụ phân loại như Hình 6.



Hình 6. Kiến trúc BERT_base [6]

V. THỰC NGHIỆM

A. Dữ liệu thực nghiệm

Bộ dữ liệu UPT được thu thập và tiền xử lý. Tổng số 6000 câu bình luận được gán nhãn: tiêu cực, trung tính và tích cực được mô tả trong Bảng 2.

Bảng 2. Bảng tỷ lệ các câu văn trong bộ dữ liệu UPT

	Tiêu cực	**Trung tính**	**Tích cực**	**Tổng**
Số lượng câu văn	2128	1760	2112	6000
Tỷ lệ (%)	35,47%	29,33%	35,20%	100%
Độ dài trung bình của câu văn	21,73	10,66	19,62	17,73

Chúng tôi tiến hành chia tập dữ liệu 6000 câu văn thành 3 tập huấn luyện, phát triển và kiểm thử mô hình 7/1/2 tương ứng: tập huấn luyện (train): 70%, tập phát triển (validation): 10% và tập kiểm tra (test): 20%.

Trong bảng 2, nhãn trung tính có tỷ lệ thấp hơn 2 nhãn tiêu cực và tích cực khoảng 6% nhưng không đáng kể, dữ liệu tương đối cân bằng giữa các nhãn. Độ dài trung bình của câu văn trên nhãn trung tính thấp hơn khoảng 45% so với độ dài câu của 2 nhãn tích cực và tiêu cực.

B. Công cụ đánh giá

Khi đánh giá mô hình, phép đo Precision, Recall, F1_score thường được sử dụng cho bài toán phân loại.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 - score = 2 \cdot \frac{\text{Precision. Recall}}{\text{Precision} + \text{Recall}}$$

- True Positive (TP): số lượng điểm của lớp positive được phân loại đúng là positive.
- True Negative (TN): số lượng điểm của lớp negative được phân loại đúng là negative.
- False Positive (FP): số lượng điểm của lớp negative bị phân loại nhầm thành positive.
- False Negative (FN): số lượng điểm của lớp positiv bị phân loại nhầm thành negative

Macro-average precision là trung bình cộng của các precision theo lớp, tương tự với Macro-average recall. Đối với bài toán này chúng tôi lấy giá trị trung bình cộng F1-score dựa trên Macro-average precision và Macro-average recall.

C. Cài đặt và kết quả thực nghiệm

Chúng tôi thực hiện cài đặt trên môi trường Google Colab với GPU Tesla V100 16GB, ngôn ngữ lập trình Python, thư viện Huggingface, framework Pytorch.

Chúng tôi sử dụng thuật toán tối ưu Adam cho mô hình CNN/LSTM, trong khi đó BERT-base, PhoBERT sử dụng thuật toán tối ưu AdamW và các siêu tham số trong bảng 3.

Bảng 3. Bộ siêu tham số của các mô hình CNN/LSTM/BERT-base/PhoBERT

Model	Embedding size	Learning rate	Batch size	Max-length	Epochs
CNN	128	0,001	32		10
LSTM	128	0,001	32		10
BERT-base [6]		2e-5	32	125	10
PhoBERT		2e-5	32	125	10

1. Kết quả thực nghiệm

Bảng 4 thể hiện kết quả thử nghiệm của các mô hình LSTM, CNN, BERT-base và PhoBERT.

Bảng 4. Kết quả thử nghiệm trên bộ dữ liệu UPT

Mô hình	Precision (%)	Recall (%)	F1-score (%)
LSTM	72,23	72,03	72,11
CNN	80,89	79,70	79,60
BERT-base [6]	86,71	86,18	86,25
PhoBERT	**89,89**	**89,71**	**89,68**

2. Phân tích kết quả thực nghiệm

Theo kết quả Bảng 4, các mô hình LSTM đạt F1-score 72,11%, trong khi đó CNN đạt kết quả F1-score 79,60% tốt hơn mô hình LSTM trên bộ dữ liệu ĐHPT. Chúng ta nhận thấy CNN không chỉ xử lý tốt trên nhận dạng hình ảnh mà còn áp dụng vào mảng xử lý ngôn ngữ tự nhiên rất hiệu quả.

Kết quả trên các mô hình BERT cải thiện đáng kể, cao hơn 10% so với các mô hình CNN, LSTM trước đó. Ta thấy kết quả Bảng 4, PhoBERT đạt F1-score 89,68% cao hơn mô CNN 10% và LSTM 17%. Điều này hiển nhiên, PhoBERT được huấn luyện trên bộ dữ liệu hoàn toàn bằng tiếng Việt và được tách từ trước khi đưa vào huấn luyện, chúng học các véctơ nhúng của từ theo ngữ cảnh thay vì véctơ cố định như véctơ embedding, biểu diễn được nhiều ý nghĩa khác nhau của một từ. So với BERT-base [6] thì mô hình PhoBERT cao hơn khoảng 3% nhưng không đáng kể. Một việc nữa đó là huấn luyện mô hình tinh chỉnh sẽ mất ít thời gian hơn, so với PhoBERT với epoch thứ 4 thì kết quả trên tập phát triển đạt độ chính xác accuracy: 92,71%, F1-score: 91,92% đồng nghĩa với những khuyến nghị của tác giả khi tinh chỉnh BERT chỉ từ 2-4 epoch cho một nhiệm vụ xử lý ngôn ngữ tự nhiên.

VI. KẾT LUẬN

Trong bài báo này, chúng tôi đã thu thập được bộ dữ liệu dành cho bài toán phân tích quan điểm người dùng nói chung và mảng giáo dục nói riêng. Đặc biệt, thu thập các bình luận, góp ý về Trường Đại học Phan Thiết dưa trên các

nguồn khác nhau, không chỉ là kênh sinh viên mà còn có doanh nghiệp, người dùng xã hội từ năm 2015 đến 2019. Từ việc thu thập dữ liệu, phân tích dữ liệu, các phản hồi mang tính góp ý, mặt chưa tốt được gán nhãn tiêu cực cao hơn các hai nhãn còn lại giúp cho nhà trường mảng nào cần phải cấp thiết cải thiện, phát triển, nâng cao chất lượng giảng dạy.

Bài báo đã đề xuất mô hình tiền huấn luyện PhoBERT giải quyết tốt trên bộ dữ liệu phân tích quan điểm, kết quả trên bộ dữ liệu tách từ điểm F1-score 89,68% và độ chính xác 89,89% vượt trội so với các mô hình trước khi BERT ra đời CNN, LSTM. Kết quả cao hơn mô hình tiền huấn luyện BERT-base được đề xuất trong bài báo của nhóm tác giả Quoc Thai Nguyen và cộng sự.

TÀI LIỆU THAM KHẢO

[1] Mindez, Jose, Lobel, Hans, Parra, Denis, Herrera, Juan, Using twitter to infer user satisfaction with public transport: the case of Santiago, Chile. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2915107, 2019.

[2] Trinh, Son & Nguyen, Luu & Vo, Minh, Combining lexicon-based and learning-based methods for sentiment analysis for Product Reviews in Vietnamese language. 10.1007/978-3-319-60170-0_5, 2018.

[3] Nguyen, Hong Nam & Le, Thanh & Le, Hai & Pham, Tran Vu, Domain specific sentiment dictionary for opinion mining of Vietnamese text. 10.1007/978-3-319-13365-2_13, 2014.

[4] Tran, Bang & Haruechaiyasak, Choochart & Sornlertlamvanich, Virach, Vietnamese sentiment analysis based on term feature selection approach, 2015.

[5] Devlin, J., Chang, M., Lee, K., & Toutanova, K., BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT, 2019.

[6] Q. T. Nguyen, T. L. Nguyen, N. H. Luong and Q. H. Ngo, "Fine-Tuning BERT for sentiment analysis of vietnamese reviews", 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), pp. 302-307, doi: 10.1109/NICS51282.2020.9335899, 2020.

[7] Vo, Hung & Lam, Hai & Nguyen, Duc Dung & Tuong, Nguyen. Topic classification and sentiment analysis for Vietnamese education survey system. Asian Journal of Computer Science and Information Technology. 6. 27-34. 10.15520/ajcsit.v6i3.44.g31, 2016.

[8] Nguyen, Kiet & Nguyen, Vu & Nguyen, Phu & Truong, Tham & Nguyen, Ngan. UIT-VSFC: Vietnamese students' feedback corpus for sentiment analysis. 19-24. 10.1109/KSE.2018.8573337, 2018.

[9] Nguyen, Phu & Hong, Tham & Nguyen, Kiet & Nguyen, Ngan. Deep learning versus traditional classifiers on vietnamese students' feedback corpus. 75-80. 10.1109/NICS.2018.860683, 2018.

[10] Cummins, Stephen & Burd, Liz & Hatch, Andrew. Using feedback tags and sentiment analysis to generate sharable learning resources investigating automated sentiment analysis of feedback tags in a Programming Course. 653-657. 10.1109/ICALT.2010.186, 2010.

[11] Menaha, R., Dhanaranjani, R., Rajalakshmi, T., & Yogarubini, R. Student feedback mining system using sentiment analysis. IJCATR, 6, 1-69, 2017.

[12] Tummel, A. A. D. J. C., & Richert, S. J. A. (2015, December). Sentiment analysis of social media for evaluating universities. In the second International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2015) (p. 49).

[13] Baradwaj, B. K., & Pal, S. Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417, 2012.

[14] Vo, H. T., Lam, H. C., Nguyen, D. D., & Tuong, N. H. Topic classification and sentiment analysis for Vietnamese education survey system. Asian Journal of Computer Science and Information Technology, 6(3), 27-34, 2016.

[15] Kim, Y. (2014). Convolutional neural networks for sentence classification., arXiv. preprint.

[16] Hochreiter, S., & Schmidhuber, J. Long short-term memory. Neural computation, 9(8), 1735-1780, 1997.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,... & Polosukhin, I., Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.

[18] Nguyen, D. Q., & Nguyen, A. T., PhoBERT: Pre-trained language models for Vietnamese. arXiv preprint arXiv:2003.00744, 2020.

ANALYSIS OF SOCIAL PERSPECTIVES ON PHAN THIET UNIVERSITY

Vo Quoc Tuan, Tran Thanh Phuoc, Tran Thanh Tram

ABSTRACT: *In modern education, universities is service providers, students and parents are customers. It is extremely necessary to capture the emotions and views of these customers (also known as society) about the services schools have provided. Through those feedbacks, schools will receive positive and negative evaluations to promote the positive sides and limit the negative sides. Within the scope of this article, we focus on two things: (1) Building the data set of social sentiment for Phan Thiet University including 3 labels: positive, neutral and negative; (2) Proposing to use deep learning models such as CNN, LSTM, BERT and PhoBERT to test the case of perspectives analysis of Phan Thiet University. The results show that PhoBERT gave the highest results on the Phan Thiet University dataset with an F1-score of 89.68%.*