SỬ DỤNG BERT VÀ CÂU PHỤ TRỢ CHO TRÍCH XUẤT KHÍA CẠNH TRONG VĂN BẢN TIẾNG VIỆT

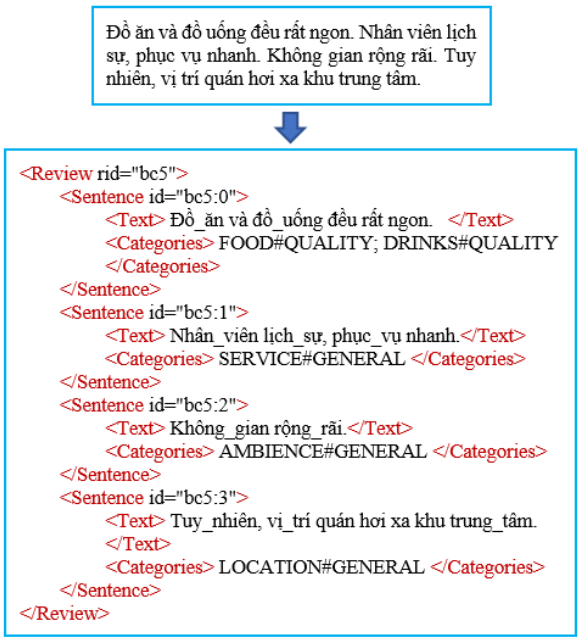
Nguyễn Ngọc Điệp, Nguyễn Thị Thanh Thủy
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Trích xuất khía cạnh (aspect extraction) là một nhiệm vụ trong bài toán khai phá quan điểm dựa trên khía cạnh (aspect-based opinion mining), nhằm xác định và phân loại các cụm từ quan điểm (opinion target) về những đặc tính của sản phẩm trong văn bản có thể hiện quan điểm. Đa phần các nghiên cứu trước về trích xuất khía cạnh và khai phá quan điểm dựa trên khía cạnh là cho văn bản tiếng Anh, có rất ít nghiên cứu cho tiếng Việt. Các nghiên cứu cho tiếng Việt có độ chính xác cao hơn thường dựa trên các phương pháp học có giám sát hoặc dựa trên học sâu, với các mô hình phụ thuộc vào nhúng từ độc lập ngữ cảnh (như word2vec). Bài báo này trình bày một phương pháp trích xuất khía cạnh dựa trên khả năng mô hình hóa với nhúng từ theo ngữ cảnh, sử dụng các mô hình ngôn ngữ được huấn luyện sẵn như BERT. Khác với các nghiên cứu trước đó sử dụng một câu dữ liệu đầu vào rồi sau đó trích xuất ra các khía cạnh có trong câu, bài báo đề xuất sử dụng câu phụ trợ được tạo ra từ các từ khóa khía cạnh nhằm tận dụng được thông tin quan trọng đã biết, kết hợp với câu đầu vào ban đầu để tạo ra cặp câu đầu vào cho BERT. Mô hình đề xuất dựa trên BERT có thêm một lớp tuyến tính để phân loại, được tinh chỉnh cùng với câu phụ trợ cho thấy kết quả rất tốt trên kho ngữ liệu có sẵn, đã chú thích về các loại danh mục khía cạnh (aspect category) được thu thập từ những bài đánh giá/bình luận về nhà hàng trên mạng xã hội bằng ngôn ngữ tiếng Việt.

Từ khóa: trích xuất danh mục khía cạnh, khai phá quan điểm cho tiếng Việt, mô hình ngôn ngữ huấn luyện sẵn, mô hình BERT

I. GIỚI THIỆU

Khai phá quan điểm dựa trên khía cạnh [1], [2], [3], [4] đã trở thành một chủ đề nghiên cứu hấp dẫn trong cả hai lĩnh vực xử lý ngôn ngữ tự nhiên và khai phá dữ liệu trong những năm gần đây. Khai phá quan điểm dựa trên khía cạnh nhằm xác định cảm xúc cho từng khía cạnh của sản phẩm/dịch vụ được diễn tả trong văn bản, thay vì chỉ xác định mỗi cảm xúc (tích cực, tiêu cực hay trung tính) trên toàn bộ văn bản. Có hai nhiệm vụ trong bài toán khai phá quan điểm dựa trên khía cạnh, đó là: 1) Trích xuất khía cạnh, trong đó xác định các loại danh mục khía cạnh (các cặp thực thể và thuộc tính) được diễn tả ý kiến/quan điểm trong văn bản; và 2) Phân loại cảm xúc, trong đó gán một nhãn phân cực (tích cực, tiêu cực hoặc trung tính) cho mỗi khía cạnh đã được xác định. Trong bài báo này, chúng tôi chỉ tập trung vào nhiệm vụ thứ nhất. Đây là một nhiệm vụ khó khăn vì người dùng thường sử dụng các từ các khác nhau để diễn tả về cùng một khía cạnh hoặc thậm chí chỉ ngụ ý đề cập đến khía cạnh. Để minh họa cho các loại danh mục khía cạnh, ta xem xét một bình luận của khách hàng về một nhà hàng được chú thích như trong Hình 1 dưới đây.



Hình 1. Ví dụ về một bài bình luận/đánh giá với các danh mục khía cạnh được chú thích

Bài bình luận gồm có 4 câu, trong đó câu thứ nhất thuộc 2 nhãn khía cạnh là FOOD#QUALITY (chất lượng đồ ăn) và DRINKS#QUALITY (chất lượng đồ uống), câu thứ hai thuộc 1 nhãn SERVICE#GENERAL (dịch vụ nhà hàng), câu thứ ba thuộc 1 nhãn AMBIENCE#GENERAL (không gian nhà hàng), và câu thứ tư thuộc 1 nhãn LOCATION#GENERAL (vị trí nhà hàng).

Các công trình nghiên cứu ban đầu về trích xuất khía cạnh theo cách tiếp cận dựa trên học có giám sát [1], [2], [5], [6], [7], [8], [9], [10], [11], thường sử dụng phương pháp trích chọn đặc trưng thủ công từ tập văn bản đầu vào và thực hiện xác định các khía cạnh bằng các bộ phân lớp.

Tác giả liên hệ: Nguyễn Thị Thanh Thủy,
Email: thuyr205@gmail.com
Đến tòa soạn: 9/2022, chỉnh sửa: 10/2022, chấp nhận đăng: 11/2022.

Các phương pháp này khá hiệu quả và đạt được độ chính xác cao, đã được chứng minh trong nhiều nghiên cứu trước đây [1], [12]. Tuy nhiên, để trích chọn được các đặc trưng thủ công tốt, cần phụ thuộc vào kinh nghiệm và kiến thức của các chuyên gia trong nhiều miền lĩnh vực khác nhau, do vậy đây là một thách thức lớn.

Một cách tiếp cận tiên tiến hơn là sử dụng nhúng từ (word embeddings) với các mô hình học sâu. Nhúng từ không theo ngữ cảnh (như word2vec, fastText) chỉ bao gồm một tập hợp các nhúng từ tĩnh và không xem xét đến các ngữ cảnh khác nhau mà chúng có thể xuất hiện, do đó bị hạn chế về mức độ hiệu quả. Ngược lại, các mô hình ngôn ngữ được huấn luyện trước như ELMO [14] hay BERT [13] có thể tạo ra các véc-tơ nhúng từ động, có thể thay đổi khi ngữ cảnh của các từ trong câu thay đổi. Điều này làm cho việc sử dụng BERT trong nhiều tác vụ được khuyến khích hơn, thông qua việc tinh chỉnh mới trên tập dữ liệu liên quan đến từng nhiệm vụ và miền lĩnh vực cụ thể.

Bài báo này sử dụng BERT làm mô hình cơ sở để cải thiện các mô hình trích xuất danh mục khía cạnh, cho phép sử dụng các tài nguyên bổ sung như dữ liệu huấn luyện phụ trợ bên ngoài. Việc này dựa trên cơ sở là biểu diễn đầu vào của BERT có thể đại diện cho cả một câu đơn lẻ hoặc một cặp câu. Do đó, có thể chuyển đổi nhiệm vụ trích xuất danh mục khía cạnh thành một nhiệm vụ phân loại cặp câu đầu vào gồm câu ban đầu và câu phụ trợ. Trong nghiên cứu này, câu phụ trợ được xây dựng dựa trên nhãn của khía cạnh trong danh mục. Đồng thời, kết hợp với việc tinh chỉnh mô hình BERT được huấn luyện trước, mô hình đề xuất đạt được kết quả tốt hơn so với nghiên cứu trước đó [27]. Ngoài ra, do tiếng Việt là một ngôn ngữ nghèo tài nguyên nên rất ít mô hình BERT được huấn luyện trước cho các nhiệm vụ cụ thể. Bài báo sẽ thực hiện các thử nghiệm trên mô hình BERT đa ngôn ngữ và so sánh với mô hình PhoBERT huấn luyện trước cho tiếng Việt và tinh chỉnh cho bài toán phân loại cặp câu.

Đóng góp của bài báo bao gồm ba phần:

- Thứ nhất, đề xuất giải pháp trích xuất khía cạnh cho văn bản tiếng Việt dựa trên BERT, thông qua giải quyết bài toán phân loại cặp câu đầu vào. Mô hình đề xuất sử dụng câu phụ trợ được tạo ra từ các nhãn trong danh mục khía cạnh. Mô hình này có khả năng tự học đặc trưng (không cần kiến thức chuyên gia để trích chọn đặc trưng thủ công).
- Thứ hai, tinh chỉnh mô hình huấn luyện BERT để đạt được hiệu quả cao hơn trên tập dữ liệu văn bản tiếng Việt đã được gán nhãn.
- Thứ ba, so sánh và phân tích hiệu quả của mô hình đề xuất khi sử dụng mô hình BERT huấn luyện sẵn đa ngôn ngữ và PhoBERT.

Phần còn lại của bài báo được tổ chức như sau. Phần II mô tả cơ sở lý thuyết. Phần III trình bày các nghiên cứu liên quan. Phần IV đề xuất phương pháp thực hiện trích xuất danh mục khía cạnh trong văn bản tiếng Việt. Chi tiết về bộ dữ liệu và các thực nghiệm được trình bày trong phần Phần V và Phần VI. Phần VII là kết luận bài báo và định hướng nghiên cứu.

II. CƠ SỞ LÝ THUYẾT

A. BERT

BERT (Bidirectional Encoder Representation from Transformer) là mô hình biểu diễn từ theo 2 chiều, sử dụng kỹ thuật Transformer [13]. BERT được thiết kế để huấn luyện trước các biểu diễn từ nhúng (pre-train word embedding). Điểm đặc biệt ở BERT đó là có thể điều hòa cân bằng ngữ cảnh theo cả 2 chiều trái và phải. Không như những kỹ thuật khác, BERT không sử dụng LSTM (Long short-term memory) để trích xuất đặc trưng ngữ cảnh của từ, mà thay vào đó sử dụng Transformer, là cơ chế dựa trên sự tập trung (attention) chứ không dựa trên sự lặp lại. So với các mô hình dựa trên huấn luyện theo 2 chiều riêng biệt trái, phải như ELMO [14], BERT có thể trích xuất nhiều đặc trưng ngữ cảnh hơn từ cùng một văn bản. Nhờ vào khả năng xử lý một cách hiệu quả sự không tường minh (ngữ nghĩa tiềm ẩn), và đây chính là điểm khác biệt nhất của việc hiểu ngôn ngữ tự nhiên, BERT có thể đạt độ chính xác cao trong việc phân tích các ngôn ngữ gần gũi với con người. Các nghiên cứu đã cho thấy, nhờ sử dụng một lượng lớn dữ liệu văn bản huấn luyện, BERT đã đạt được kết quả tối ưu tốt nhất trên 11 tác vụ xử lý ngôn ngữ tự nhiên (NLP) [13].

Quá trình huấn luyện trước bên trái và bên phải của BERT đạt được bằng cách sử dụng mặt nạ mô hình ngôn ngữ đã được sửa đổi, gọi là masked language model (MLM). Mục đích của MLM là che giấu một từ ngẫu nhiên trong một câu với xác suất nhỏ. Khi mô hình che một từ, nó sẽ thay thế từ đó bằng một token [MASK]. Sau đó, mô hình cố gắng dự đoán từ bị che bằng cách sử dụng ngữ cảnh của cả bên trái và phải của từ bị che, với sự hỗ trợ của transformer. Ngoài việc trích xuất ngữ cảnh trái và phải bằng MLM, BERT có thêm một mục tiêu quan trọng khác so với các nghiên cứu trước, đó là dự đoán câu tiếp theo (Next Sentence Prediction).

Biểu diễn đầu vào

Văn bản đầu vào của mô hình BERT trước tiên được xử lý thông qua một phương pháp được gọi là tách từ wordpiece [30]. Kết quả sẽ tạo ra một tập các token, mỗi token đại diện cho một từ. Chuỗi đầu vào BERT biểu diễn một cách tường minh cho cả dạng văn bản đơn và cặp văn bản. Với văn bản đơn, chuỗi đầu vào BERT là sự ghép nối của token đặc biệt "<CLS>", token của chuỗi văn bản, và token phân tách "<SEP>". Trạng thái ẩn cuối cùng tương ứng với token phân loại [CLS] được sử dụng làm véc-tơ biểu diễn tổng hợp cho các nhiệm vụ phân loại [13]. Với cặp văn bản, chuỗi đầu vào BERT là sự ghép nối của "<CLS>", token của chuỗi văn bản đầu, "<SEP>", token của chuỗi văn bản thứ hai, và "<SEP>". Nhờ vậy, BERT có thể được sử dụng để so sánh một cặp hai câu. Tập hợp các token của cặp câu này sau đó được xử lý thông qua ba lớp nhúng khác nhau có cùng kích thước, rồi được cộng lại với nhau và chuyển đến lớp mã hóa (encoder layer). Ba lớp nhúng đó là: lớp nhúng token (token embedding layer), lớp

nhúng đoạn (segment embedding layer) và lớp nhúng vị trí (position embedding layer).

Transformer

Các nghiên cứu về mô hình hóa chuỗi trước đây sử dụng một kiểu kiến trúc chung là seq2seq [31], dựa trên các kỹ thuật như RNN [32] và LSTM [33]. Kiến trúc transformer không dựa trên RNN mà dựa trên kỹ thuật tập trung (attention) [34]. Nó quyết định những chuỗi nào là quan trọng trong mỗi bước tính toán. Bộ mã hóa không chỉ ánh xạ đầu vào thành véc-tơ không gian nhiều chiều hơn mà còn sử dụng các từ khóa quan trọng làm đầu vào bổ sung cho bộ giải mã. Nhờ vậy có thể cải thiện bộ giải mã vì có thêm thông tin bổ sung về chuỗi quan trọng và từ khóa cung cấp ngữ cảnh cho câu.

Nhiệm vụ phân loại cặp câu

Ban đầu, BERT huấn luyện trước mô hình nhằm có được các véc-tơ nhúng từ để tinh chỉnh mô hình phù hợp hơn cho một nhiệm vụ cụ thể, mà không cần phải thay đổi nhiều về mô hình và tham số. Thông thường, chỉ cần bổ sung thêm một lớp đầu ra trên đỉnh của mô hình để làm cho nó có thể dễ dàng đáp ứng hơn trong một nhiệm vụ cụ thể. Nhiệm vụ phân loại cặp câu đề cập đến việc xác định các quan hệ ngữ nghĩa giữa hai câu. Mô hình lấy hai văn bản làm đầu vào và xuất ra một nhãn đại diện cho kiểu quan hệ giữa các câu. Loại nhiệm vụ này đánh giá mô hình về mức độ hiểu biết ngôn ngữ tự nhiên và khả năng suy luận sâu hơn về các câu đầy đủ [35].

Mô hình ngôn ngữ

Việc xây dựng một mô hình ngôn ngữ thông thường cần sử dụng một lượng lớn dữ liệu văn bản không chú thích, được gọi là huấn luyện trước (pre-training). Mục đích chung của mô hình ngôn ngữ là học biểu diễn theo ngữ cảnh của từ. Mô hình ngôn ngữ là thành phần quan trọng trong việc giải quyết các bài toán NLP, tìm hiểu sự xuất hiện của từ và các mẫu dự đoán từ dựa trên dữ liệu văn bản không có chú thích. Mô hình ngôn ngữ học ngữ cảnh bằng cách sử dụng các kỹ thuật như nhúng từ, tức là sử dụng véc-tơ để đại diện cho các từ trong không gian véc-tơ [29]. Dựa trên lượng lớn dữ liệu huấn luyện, mô hình ngôn ngữ học các biểu diễn của các từ tùy thuộc vào ngữ cảnh, nhờ đó cho phép các từ tương tự nhau có biểu diễn tương tự.

Masked Language Model

BERT sử dụng token mặt nạ [MASK] để huấn luyện trước các biểu diễn sâu hai chiều cho mô hình ngôn ngữ. Trái ngược với các mô hình ngôn ngữ khác thực hiện huấn luyện từ trái sang phải hoặc từ phải sang trái để dự đoán các từ, và từ được dự đoán được đặt ở cuối hoặc ở đầu chuỗi văn bản, thì BERT che dấu một từ ngẫu nhiên trong chuỗi.

Dự đoán câu tiếp theo (NSP)

Dự đoán câu tiếp theo được sử dụng để hiểu mối quan hệ giữa hai câu văn bản. BERT đã được huấn luyện trước để dự đoán liệu có tồn tại mối quan hệ giữa hai câu hay không. Mỗi câu có kích thước nhúng riêng.

B. RoBERTa

RoBERTa [28] là một tiếp cận kế thừa kiến trúc và thuật toán của mô hình BERT nhưng mạnh và tối ưu hơn. Dự án này của Facebook hỗ trợ việc huấn luyện lại các mô hình BERT trên những bộ dữ liệu mới cho các ngôn ngữ khác ngoài một số ngôn ngữ phổ biến. Hiện đã có rất nhiều các mô hình huấn luyện trước cho những ngôn ngữ khác nhau được huấn luyện trên RoBERTa, trong đó có tiếng Việt (mô hình PhoBERT).

RoBERTa lặp lại các thủ tục huấn luyện như mô hình BERT, nhưng có một thay đổi đó là huấn luyện mô hình lâu hơn, với batch size lớn hơn và trên nhiều dữ liệu hơn. Ngoài ra, để nâng cao độ chính xác trong biểu diễn từ thì RoBERTa đã loại bỏ tác vụ dự đoán câu tiếp theo và huấn luyện trên các câu dài hơn. Đồng thời, mô hình cũng thay đổi kiểu che giấu (masking) một cách linh hoạt hơn cho dữ liệu huấn luyện, bằng cách ẩn đi một số từ ở câu đầu ra bằng token mặt nạ.

Tuy loại bỏ tác vụ dự đoán câu tiếp theo trong khi huấn luyện, RoBERTa vẫn nhận dữ liệu đầu vào theo cả 2 cách là câu đơn và cặp câu như mô hình BERT. Đồng thời, kết quả trên các tác vụ sử dụng các cặp câu đầu vào vẫn vượt trội hơn so với mô hình BERT [28].

C. Mô hình ngôn ngữ huấn luyện trước dựa trên BERT cho tiếng Việt

Hai mô hình huấn luyện trước phổ biến cho tiếng Việt gồm mô hình BERT đa ngôn ngữ [13] và PhoBERT [21]. Trong đó, mô hình BERT đa ngôn ngữ được huấn luyện trên 104 ngôn ngữ, là mô hình BERT huấn luyện sẵn duy nhất cho hầu hết các ngôn ngữ. Mô hình BERT đa ngôn ngữ đã cho thấy hiệu quả trong hầu hết các nhiệm vụ xử lý ngôn ngữ tự nhiên đa ngôn ngữ như phân tích cảm xúc [22], [23] phân loại văn bản [24] sinh văn bản [25] và hội thoại [26].

PhoBERT là một mô hình đơn ngôn ngữ cho tiếng Việt. PhoBERT dựa trên kiến trúc RoBERTa [28] và cho thấy hiệu suất tốt hơn hẳn so với các phương pháp dựa trên mô hình BERT đa ngôn ngữ khi làm việc với văn bản tiếng Việt. Trong nghiên cứu [21], tác giả đã chỉ ra sự ưu việt của PhoBERT trên 4 tác vụ xử lý ngôn ngữ tự nhiên, gồm gán nhãn từ loại (POS tagging), phân tích sự phụ thuộc về cú pháp (Dependency parsing), nhận diện thực thể trong câu (NER) và suy luận ngôn ngữ (NLI).

III. CÁC NGHIÊN CỨU LIÊN QUAN

Phần này trình bày ngắn gọn về các nghiên cứu liên quan đến trích xuất danh mục khía cạnh và các nghiên cứu liên quan được xử lý cho ngôn ngữ tiếng Việt.

Trích xuất danh mục khía cạnh là nhiệm vụ con đầu tiên trong khai thác quan điểm dựa trên khía cạnh. Các phương pháp hiện có cho trích xuất danh mục khía cạnh có thể được chia thành hai cách tiếp cận chính, dựa trên phân loại và dựa trên phân cụm. Đối với cách tiếp cận dựa trên phân cụm, các thuật toán học không giám sát hoặc bán giám sát với kỹ thuật lấy mẫu có thay thế (bootstrapping) [2], [5], [9] là một lựa chọn thích hợp do không có dữ liệu đã được

chú thích sẵn. Các phương pháp sử dụng cách tiếp cận này thường trích xuất các loại khía cạnh theo hai bước: 1) trích xuất tất cả các thuật ngữ khía cạnh từ một văn bản được cho; và 2) phân cụm các thuật ngữ khía cạnh có ý nghĩa tương tự nhau thành các nhóm danh mục khía cạnh.

Các phương pháp cho cách tiếp cận dựa trên phân loại sử dụng các thuật toán học có giám sát và thường mô hình hóa nhiệm vụ như một bài toán phân loại đa nhãn trong đó mỗi nhãn tương ứng với một danh mục khía cạnh, hoặc nhiều bài toán phân loại nhị phân trong đó mỗi bài toán xử lý một loại khía cạnh cụ thể [1], [2], [5], [6], [7], [8], [9], [10], [11], [12], [15], [16]. So với cách tiếp cận dựa trên phân cụm, cách tiếp cận dựa trên phân loại chính xác hơn bởi nó có thể sử dụng học có giám sát với dữ liệu đã được chú thích. Nghiên cứu của chúng tôi ở đây thuộc cách tiếp cận dựa trên phân loại, trong đó chúng tôi sử dụng các kỹ thuật học sâu dựa theo kiến trúc BERT trích xuất danh mục khía cạnh. Mô hình được tinh chỉnh bằng cách sử dụng thêm các câu phụ trợ nhằm đưa về bài toán phân loại cặp câu. Phương pháp sử dụng một lớp tuyến tính để phân loại danh mục khía cạnh. Cách tiếp cận này tương tự như cách tiếp cận của nhóm tác giả trong [36], đó là chuyển bài toán phân loại cảm xúc theo khía cạnh thành bài toán phân loại một cặp câu, cụ thể là trả lời câu hỏi hoặc suy luận ngôn ngữ. Tuy nhiên, cách tiếp cận này chưa được sử dụng trong các bài toán phân tích khía cạnh cho ngôn ngữ tiếng Việt.

Trong số các nghiên cứu trước đây về khai thác quan điểm dựa trên khía cạnh tiếng Việt, tác giả Vu và cộng sự [17] trình bày một phương pháp dựa trên luật để khai thác đánh giá sản phẩm. Trong đó, từ thể hiện khía cạnh và từ thể hiện quan điểm được trích xuất bằng cách sử dụng một bộ luật cú pháp tiếng Việt. Tác giả Le và cộng sự [18] trình bày một phương pháp bán giám sát để trích xuất và phân loại các thuật ngữ khía cạnh trong văn bản tiếng Việt. Phương pháp của họ có thể được phân vào nhóm phương pháp dựa trên phân cụm. Gần đây, một số nghiên cứu đã được trình bày sử dụng phương pháp học có giám sát [19], [20] cho bài toán kết hợp phân tích cảm xúc dựa trên khía cạnh tiếng Việt tại hội thảo quốc tế lần thứ năm về Xử lý ngôn ngữ và tiếng Việt (VLSP 2018) [19].

Nghiên cứu của chúng tôi khác với những nghiên cứu trước đây ở chỗ chúng tôi đề xuất một phương pháp mới dựa trên kiến trúc BERT cùng với câu phụ trợ để giải quyết nhiệm vụ trích xuất danh mục khía cạnh trong tiếng Việt. Dựa trên cặp câu đầu vào, nhiệm vụ này được chuyển thành nhiệm phụ phân loại/xác định mối quan hệ của cặp câu đầu vào. Thêm nữa, nghiên cứu xác định các loại danh mục khía cạnh trong một câu thay vì toàn bộ bài bình luận/đánh giá. Giải quyết nhiệm vụ ở cấp độ câu là thực tế hơn và có thể áp dụng tốt hơn trong thế giới thực.

IV. PHƯƠNG PHÁP ĐỀ XUẤT

Phần này trình bày chi tiết về phương pháp đề xuất trích xuất danh mục khía cạnh trong văn bản tiếng Việt.

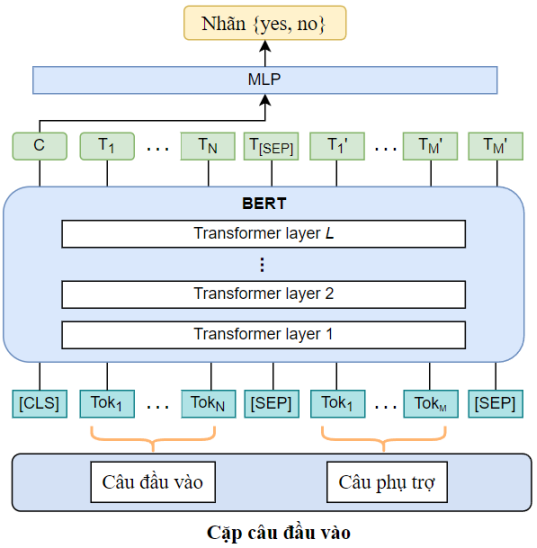
A. Mô tả bài toán

Cho một câu đầu vào *s* trong tập văn bản gồm một chuỗi các từ $\{w_1, \dots, w_m\}$. Giả thiết có một tập cố định các loại danh mục khía cạnh A = $\{a_1, \dots, a_k\}$. Nhiệm vụ là cần thực hiện trích xuất tập danh mục khía cạnh {(a): a ∈ A} có trong câu *s*. Trong tập dữ liệu, mỗi khía cạnh thường được mô tả dưới dạng cặp "ENTITY#ASPECT" (Thực thể/khía cạnh).

Ví dụ, với một câu đầu vào "Đồ_ăn ngon nhưng khá đắt." Tập danh mục khía cạnh trích xuất được gồm 2 khía cạnh: "FOOD#QUALITY" và "FOOD#PRICES".

B. Trích xuất khía cạnh sử dụng BERT

Trong nghiên cứu này, chúng tôi sử dụng kiến trúc BERT với các lớp transformer để xác định các biểu diễn ngữ cảnh tương ứng cho chuỗi token đầu vào, bao gồm cặp văn bản từ câu đánh giá và câu phụ trợ (được tạo ra từ danh mục khía cạnh). Sau đó, các biểu diễn theo ngữ cảnh này được đưa vào lớp tác vụ cụ thể để dự đoán các nhãn. Kiến trúc tổng thể của mô hình đề xuất được trình bày trong Hình 2.



Hình 2. Kiến trúc tổng thể của mô hình

1) Xây dựng câu phụ trợ

Để chuyển đổi bài toán trích xuất danh mục khía cạnh thành bài toán phân loại cặp câu đầu vào dựa trên mô hình BERT, mô hình đề xuất dựa trên cấu trúc của một bộ phân loại cho một cặp câu, với dữ liệu đầu vào là một câu đánh giá/nhận xét trong văn bản ban đầu và câu phụ trợ dựa trên một khía cạnh. Mô hình này được sử dụng để dự đoán xem liệu khía cạnh có liên quan tới câu đầu vào hay không, dựa trên nhãn "yes" và "no" tương ứng. Như vậy, bài toán ban đầu được chuyển thành bài toán phân loại nhị phân (nhãn ∈ {yes, no}). Nói cách khác, mô hình nhận câu đánh giá là câu đầu tiên và các khía cạnh là câu phụ trợ, và nhiệm vụ sẽ là xác định mối quan hệ của câu đánh giá đối với từng khía cạnh. Do trong tập dữ liệu, các khía cạnh thường có định dạng "ENTITY#ASPECT", khi đó câu phụ trợ sẽ được viết lại đơn giản là "ENTITY, ASPECT.".

2) Biểu diễn đầu vào và phương pháp hoạt động của mô hình dựa trên BERT

Đối với mô hình trích xuất danh mục khía cạnh đề xuất dựa trên BERT, đầu vào được BERT xử lý theo cách đặc biệt, trong đó một cặp câu văn bản gồm câu bình luận/đánh giá của người dùng và câu phụ trợ được thực hiện tách từ để thành một chuỗi các token. Chuỗi token này được kết hợp với các token đặc biệt gồm [CLS] và [SEP] để làm đầu vào cho mô hình. Đầu ra mô hình thể hiện mối quan hệ giữa câu đánh giá và câu phụ trợ là có quan hệ (nhãn "yes") và không quan hệ (nhãn "no").

Đầu vào: [CLS] câu bình luận/đánh giá [SEP] câu phụ trợ [SEP]

Đầu ra: yes/no

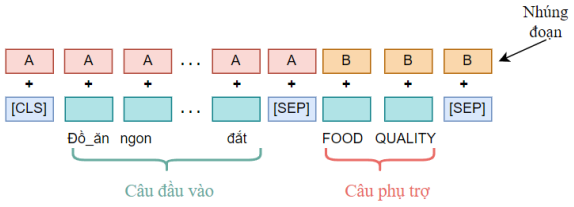
Nhúng từ/token (Token Embeddings): Các token đặc biệt gồm [CLS] được thêm vào đầu câu, [SEP] được thêm vào cuối câu bình luận/đánh giá để phân biệt với câu phụ trợ, và 1 token [SEP] được thêm vào cuối câu phụ trợ. Chuỗi token bao gồm cả 2 ký hiệu [CLS] và [SEP] trước tiên được đưa vào lớp embedding, để tạo ra các véc-tơ nhúng token (token embeddings). Các véc-tơ nhúng này không bao gồm thông tin vị trí được thêm bởi nhúng vị trí (position embeddings).

Nhúng đoạn (Segment Embeddings): để xác định xem mỗi token có liên kết với câu đầu hay câu phụ trợ hay không, một ký hiệu đánh dấu được thêm vào mỗi token trong chuỗi. Trong mô hình đề xuất, các token được đánh dấu là A thuộc về câu đánh giá và các token được đánh dấu là B thuộc về câu phụ trợ. Cách thức *nhúng đoạn* được trình bày như trong Hình 3, với ví dụ của cặp câu đầu vào gồm câu bình luận/đánh giá là "Đồ_ăn ngon nhưng khá đắt." và một câu phụ trợ mô tả một khía cạnh có trong câu là "FOOD, QUALITY."

Tiếp theo, cả véc-tơ nhúng token, véc-tơ nhúng đoạn, và véc-tơ nhúng vị trí cho mỗi token được kết hợp lại và cho vào L lớp transformer để tối ưu hóa đặc trưng ở mức token. Véc-tơ biểu diễn đầu ra H^{l-1}từ lớp transformer l - 1 được dùng làm đầu vào cho tầng transformer thứ l , trong đó $l \in [1, L]$. Véc-tơ $H^l = \{h_1^l, \dots, h_t^l\}$, tại tầng thứ l được tính như sau:

$$H_l = Transformer(H^{l-1})$$

Trong đó t là số token đầu vào. Kết quả đầu ra từ lớp transformer cuối cùng H^l được coi là véc-tơ biểu diễn theo ngữ cảnh đầy đủ cho các token đầu vào và được sử dụng làm đầu vào cho lớp tác vụ cụ thể - bộ phân lớp nhị phân.



Hình 3. Ví dụ về nhúng đoạn (Segment embeddings)

Bộ phân lớp nhị phân: Như đã chỉ ra trên Hình 2, bộ phân lớp đề xuất dựa trên BERT chỉ thêm 1 perceptron đa tầng (MLP) gồm hai tầng kết nối đầy đủ. Perceptron đa tầng này biến đổi h_t^l thành 2 đầu ra của bộ phân lớp: $P = MLP(h_t^l)$. P nhận giá trị {yes, no}, thể hiện sự liên hệ của câu đánh giá và câu phụ trợ chứa từ khóa khía cạnh, nghĩa là có khía cạnh đó hay không. Trong thực nghiệm, chúng tôi sử dụng véc-tơ đầu ra h_C^L , biến đổi biểu diễn BERT của token đặc biệt [CLS], là token mã hóa thông tin của cả cặp câu đánh giá và câu phụ trợ để làm đầu vào cho lớp perceptron đa tầng. Véc-tơ biểu diễn này thường được sử dụng trong các mô hình BERT cho bài toán phân loại câu.

V. TẬP DỮ LIỆU

1) Tập dữ liệu tiếng Việt

Tập dữ liệu tiếng Việt sử dụng cho các thử nghiệm nghiên cứu là tập dữ liệu đã được chú thích cho bài toán trích xuất khía cạnh [27]. Nguồn gốc tập dữ liệu này được thu thập từ trang web Foody *(https://www.foody.vn/)*. Sau khi thu được dữ liệu thô, sẽ tiến hành một số bước tiền xử lý như làm sạch dữ liệu, phát hiện ranh giới câu, tách từ tiếng Việt. Kết quả, thu được tập dữ liệu bao gồm 575 bài bình luận/đánh giá từ một số nhà hàng ở Việt Nam, với tổng cộng 3796 câu tiếng Việt. Có tổng số 12 nhãn các loại khía cạnh đại diện cho 12 bộ (thực thể, đặc tính).

Thống kê chi tiết 12 loại danh mục khía cạnh và tần suất xuất hiện của chúng trong tập dữ liệu được trình bày trong Bảng I. Có thể thấy, danh mục khía cạnh có tần số xuất hiện nhiều nhất là FOOD#QUALITY, trong khi các danh mục khía cạnh có tần số xuất hiện thấp nhất là DRINKS#STYLE_OPTIONS và DRINKS#PRICES.

2) Xây dựng tập dữ liệu với câu phụ trợ

Với mỗi câu bình luận/đánh giá trong tập dữ liệu, chúng tôi xây dựng tập các câu phụ trợ tương ứng với mỗi khía cạnh xuất hiện trong danh mục khía cạnh.

Bảng I. Danh mục khía cạnh và tần suất xuất hiện

STT	DANH MỤC KHÍA CẠNH	TẦN SUẤT
1	RESTAURANT#GENERAL	233
2	RESTAURANT#PRICES	132
3	RESTAURANT#MISCELLANEOUS	194
4	FOOD#QUALITY	1357
5	FOOD#STYLE_OPTIONS	586
6	FOOD#PRICES	207
7	DRINKS#QUALITY	307
8	DRINKS#STYLE_OPTIONS	75
9	DRINKS#PRICES	56
10	SERVICE#GENERAL	487
11	AMBIENCE#GENERAL	516
12	LOCATION#GENERAL	215
	Tổng	**4365**

Ví dụ, cho câu đầu vào: "Đồ_ăn ngon nhưng khá đắt." Câu này có 2 khía cạnh được nói đến là "FOOD#QUALITY" và "FOOD#PRICES". Chúng tôi xây dựng 2 câu phụ trợ tương ứng là "FOOD, QUALITY." và "FOOD, PRICES.". Sau đó, gán 2 nhãn tương ứng cho

2 cặp câu đầu vào và câu phụ trợ là "yes". Các cặp câu đầu vào và câu phụ trợ đối với 10 khía cạnh còn lại (trong 12 danh mục khía cạnh liệt kê trong Bảng I) sẽ được gán nhãn là "no". Như vậy, với một câu bình luận/đánh giá ban đầu sẽ sinh ra 12 cặp câu, với câu đầu tiên là câu bình luận/đánh giá và 12 câu phụ trợ tương ứng. Với phương pháp xây dựng này, tập dữ liệu có số lượng nhãn "yes" ít hơn nhiều số lượng nhãn "no". Để giảm sự mất cân bằng, chúng tôi thực hiện lấy mẫu ngẫu nhiên cho các cặp câu với tỷ lệ nhãn "yes":"no" là 1:2. Tổng số cặp câu trong tập dữ liệu sẽ có khoảng 13 nghìn câu.

3) Tập dữ liệu bổ sung bằng Tiếng Anh

Để đánh giá so sánh hiệu năng với các phương pháp trước đó, một tập dữ liệu tiếng Anh (cả dữ liệu huấn luyện và dữ liệu kiểm tra) từ SemEval-2016 Task 5 [1] được sử dụng làm nguồn dữ liệu bổ sung cho các phương pháp đó. Tập dữ liệu tiếng Anh này bao gồm 440 bài bình luận/đánh giá với 2676 câu, được dịch sang tiếng Việt bằng công cụ dịch Google Translate (https://translate.google.com/), như trình bày trong [27]. Tập dữ liệu này sau đó được sử dụng kết hợp với tập dữ liệu tiếng Việt để thực hiện các thực nghiệm nhằm so sánh kết quả.

VI. CÁC THỰC NGHIỆM VÀ KẾT QUẢ

A. *Thiết lập thực nghiệm*

Dữ liệu được chia ngẫu nhiên thành 10 phần để thực hiện kiểm tra chéo như thực hiện trong [27]. 10% của tập dữ liệu huấn luyện được sử dụng cho việc tinh chỉnh tham số khi huấn luyện mô hình. Hiệu năng của các mô hình trích xuất khía cạnh được đo theo độ chính xác (precision), độ bao phủ (recall) và độ đo F_1 (F_1 score) trên mỗi loại danh mục khía cạnh.

Lấy ví dụ với danh mục khía cạnh DRINKS#PRICES (giá cả đồ uống). Giả sử A ký hiệu cho tập các câu có danh mục khía cạnh DRINKS#PRICES được xác định bởi mô hình, và B ký hiệu cho tập các câu có danh mục khía cạnh này được gán nhãn bởi người chú thích, thì độ chính xác, độ bao phủ và độ đo F_1 cho danh mục khía cạnh DRINKS#PRICES sẽ được tính như sau (tương tự cho các loại danh mục khía cạnh khác):

$$Precision = \frac{|A \cap B|}{|A|}$$

$$Recall = \frac{|A \cap B|}{|B|}$$

và

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

B. *Lựa chọn tham số*

Nghiên cứu sử dụng mô hình PhoBERT [21] được huấn luyện sẵn cho tiếng Việt trên tập dữ liệu xấp xỉ 145 triệu câu đã tách từ. Mô hình bao gồm 12 lớp ẩn, có hidden_size là 768. Mô hình đề xuất sử dụng bộ tối ưu Adam để tinh chỉnh mô hình trên tác vụ cụ thể với learning rate là 1e-5, batch size là 16 và số epoch tối đa là 50. Dropout là 0,3. Quá trình huấn luyện sẽ dừng sớm khi giá trị mất mát không giảm trong 3 epoch liên tục.

C. *Kết quả thực nghiệm*

Mục đích xây dựng các thực nghiệm:

- Giải quyết bài toán trích xuất danh mục khía cạnh từ các câu bình luận/đánh giá, so sánh hiệu năng của phương pháp đề xuất với các phương pháp trước đó trên cùng tập dữ liệu.
- Đánh giá sự ảnh hưởng của các mô hình BERT huấn luyện sẵn cho tiếng Việt đến hiệu năng của phương pháp đề xuất.

Phần sau đây sẽ mô tả các thực nghiệm và kết quả.

1) ***So sánh hiệu năng của các phương pháp***

Các thử nghiệm đầu tiên được thực hiện nhằm so sánh hiệu năng của phương pháp đề xuất với các phương pháp khác trong nghiên cứu trước đó [27], trên cùng tập dữ liệu. Các phương pháp này tiếp cận theo hướng học máy truyền thống sử dụng bộ phân lớp SVM (Support Véc-tơ Machine) với các đặc trưng thủ công được trích xuất bao gồm: *n*-grams và đặc trưng nhúng từ (word embeddings) để giải quyết và cải tiến hiệu năng của trích xuất danh mục khía cạnh.

- *Mô hình n-gram*: Mô hình này chỉ sử dụng tập dữ liệu tiếng Việt với các đặc trưng n-grams, bao gồm 3 tập đặc trưng: 1) unigrams; 2) unigrams và bigrams; và 3) unigrams, bigrams và trigrams. Mục đích là xem xét hiệu năng của mô hình trích xuất khía cạnh được xây dựng trên tập dữ liệu tiếng Việt với các đặc trưng n-grams cơ bản. Kết quả của tập đặc trưng tốt nhất là đặc trưng kết hợp unigrams và bigrams được sử dụng để so sánh.
- *Mô hình CRL (Cross-language model, Mô hình đa ngôn ngữ):* Sử dụng cả các câu trong tập dữ liệu tiếng Việt và các câu được dịch từ tập dữ liệu tiếng Anh để huấn luyện mô hình. Tập đặc trưng tương tự như trong mô hình ngram.
- *Mô hình WEmb (Mô hình word embedding):* Mô hình này tương tự như mô hình CRL nhưng bổ sung các đặc trưng nhúng từ (word embedding). Đặc trưng nhúng từ trong mô hình là các véc-tơ từ 50 chiều được huấn luyện với Word2Vec [17] (dùng phần mềm có tại https://code.google.com/archive/p/word2vec/) trên một tập dữ liệu văn bản từ Baomoi (https://baomoi.com/) với hơn 433 nghìn từ tiếng Việt.

Nghiên cứu này đề xuất sử dụng phương pháp học sâu dựa trên PhoBERT, cùng với việc áp dụng câu phụ trợ nhằm tạo thêm dữ liệu huấn luyện, không những giúp loại bỏ quá trình trích chọn đặc trưng thủ công tốn kém về thời gian và công sức mà còn mang lại hiệu quả cao hơn. Kết quả trình bày trong Bảng II cho thấy, hiệu năng trích xuất trung bình được cải thiện đáng kể. Thử nghiệm trên tập dữ liệu tiếng Việt, phương pháp đề xuất sử dụng PhoBERT (DL-PhoBERT) đạt kết quả vượt trội so với phương pháp trước, với độ đo F_1 lên tới 76,8% , tức có mức tăng xấp xỉ

6% so với phương pháp sử dụng đặc trưng *n*-gram trên cùng tập dữ liệu tiếng Việt, và hơn tới 4,5% trên độ đo F_1 so với phương pháp sử dụng sử dụng 2 loại đặc trưng là n-gram và nhúng từ trên tập dữ liệu bổ sung từ một ngôn ngữ khác (WEmb).

Bảng II. Kết quả trích xuất danh mục khía cạnh theo các phương pháp khác nhau

Phương pháp	Pre.(%)	Rec.(%)	F_1(%)
Ngram [27]	78.0	64.5	70.6
CRL [27]	81.2	65.7	71.7
Wemb [27]	80.4	67.1	72.3
DL-M-BERT	72.9	73.1	73.2
DL-PhoBERT	76.5	76.9	**76.8**

2) ***Ảnh hưởng của các mô hình BERT cho tiếng Việt***

Kết quả trình bày trong Bảng II cho thấy, cùng áp dụng phương pháp học sâu sử dụng kiến trúc BERT và câu phụ trợ dựa trên danh mục khía cạnh nhưng kể cả khi sử dụng mô hình đa ngôn ngữ BERT được huấn luyện trước (DL-M-BERT) thì vẫn có kết quả tốt hơn các phương pháp trước đó. Cụ thể, phương pháp sử dụng DL-M-BERT đạt độ đo F_1 là 73,2%, tăng thêm 0,9% so với phương pháp đạt kết quả cao nhất trong nghiên cứu trước (dùng dữ liệu bổ sung từ một ngôn ngữ khác, WEmb). Tuy nhiên, kết quả sử dụng mô hình DL-M-BERT lại kém hơn hẳn (thấp hơn 3,6%) so với mô hình đơn ngôn ngữ tiếng Việt là PhoBERT, đạt 76,8% tính theo độ đo F_1.

Kết quả này chính là bằng chứng về hiệu quả của phương pháp biểu diễn theo ngữ cảnh của BERT trong việc mã hóa các liên kết giữa các từ khóa trong danh mục khía cạnh và các từ thuộc ngữ cảnh. Đồng thời cho thấy việc mô hình đơn ngôn ngữ huấn luyện trước cho tiếng Việt PhoBERT hiệu quả hơn rất nhiều trong việc học đặc trưng ngữ cảnh của từ trong văn bản tiếng Việt.

VII. KẾT LUẬN

Bài báo đã trình bày một nghiên cứu thực nghiệm về trích xuất danh mục khía cạnh sử dụng mô hình học sâu dựa trên kiến trúc BERT. Đây là một mô hình hiện đại đã mang lại kết quả rất đáng kể trong nhiều tác vụ xử lý ngôn ngữ tự nhiên. Trong xử lý văn bản tiếng Việt, BERT cũng có thể được sử dụng hiệu quả do có một số mô hình BERT đơn, đa ngôn ngữ hỗ trợ tiếng Việt. Trong nghiên cứu này, chúng tôi đã thực hiện nhiệm vụ trích xuất danh mục khía cạnh trong văn bản tiếng Việt dựa trên mô hình BERT đa ngôn ngữ và đơn ngữ. Mô hình đề xuất có khả năng tự học các đặc trưng từ chuỗi dữ liệu văn bản đầu vào và biểu diễn hiệu quả nhờ BERT, giúp tiết kiệm được thời gian và công sức trong việc trích chọn các đặc trưng thủ công như các phương pháp học máy có giám sát truyền thống, đồng thời cho kết quả vượt trội. Mô hình BERT cho tiếng Việt (PhoBERT) đạt giá trị độ đo F_1 tốt nhất trên tập dữ liệu kiểm tra. Điều đó có nghĩa là các mô hình dành riêng cho riêng một ngôn ngữ vẫn hoạt động tốt hơn các mô hình đa ngôn ngữ trong nhiệm vụ trích xuất khía cạnh.

Trong thời gian tới, chúng tôi dự định nghiên cứu giải quyết bài toán rộng hơn, đó là trích xuất đồng thời danh mục khía cạnh và ý kiến/quan điểm về khía cạnh. Sự hiệu quả của BERT hứa hẹn mang lại kết quả tốt cho bài toán này.

TÀI LIỆU THAM KHẢO

[1] M. Pontiki et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proceedings of SemEval–2016, pp. 19–30, 2016.

[2] G. Qiu, B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. Computational linguistics, Vol. 37, No. 1, pp. 9–27, 2011.

[3] W. Wang, S.J. Pan, D. Dahlmeier, and X. Xiao. Recursive Neural Conditional Random Fields for Aspect-based Sentiment Analysis. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 616–626, 2016.

[4] Y. Zuo, J. Wu, H. Zhang, D. Wang, and K. Xu. Complementary Aspectbased Opinion Mining. IEEE Transactions on Knowledge and Data Engineering, 2018.

[5] R. He, W.S. Lee, H.T. Ng, and D. Dahlmeier. An Unsupervised Neural Attention Model for Aspect Extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 388– 397, 2017.

[6] T. Hercig, T. Brychcin, L. Svoboda, and M. Konkol. UWB at SemEval- 2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th international workshop on semantic evaluation (SemEval), pp. 342–349, 2016.

[7] Q. Liu, Z. Gao, B. Liu, and Y. Zhang. Automated Rule Selection for Aspect Extraction in Opinion Mining. In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1291–1297, 2015.

[8] T. Alvarez-Lopez, J. Juncal-Martinez, M. Fernandez-Gavilanes, E. Costa-Montenegro, and F.J. Gonzalez-Castano. GTI at SemEval-2016 Task 5: SVM and CRF for aspect detection and unsupervised aspect-based sentiment analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), pp. 306–311, 2016.

[9] A. Mukherjee and B. Liu. Aspect Extraction Through Semi-supervised Modeling. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 339–348, 2012.

[10] S. Poria, E. Cambria, and A. Gelbukh. Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network. Knowledge-Based Systems, Vol. 108, pp. 42–49, 2016.

[11] L. Shu, H. Xu, B. Liu. Lifelong Learning CRF for Supervised Aspect Extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 148–154, 2017.

[12] N. Jihan, Y. Senarath, D. Tennekoon, M. Wickramarathne, and S. Ranathunga. Multi-Domain Aspect Extraction using Support Véc-tơ Machines. In Proceedings of the Conference on Computational Linguistics and Speech Processing (ROCLING), pp. 308–322, 2017.

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

[14] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. arXiv:1802.05365 [cs]. ArXiv: 1802.05365. 2018.

[15] N. Jihan, Y. Senarath, D. Tennekoon, M. Wickramarathne, and S. Ranathunga, "Multi-Domain Aspect Extraction using Support Véc-tơ Machines", ROCLING 2017, pp. 308–322.

[16] D. Xenos, P. Theodorakakos, J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "AUEB-ABSA at SemEval-2016 Task 5: Ensembles of Classifiers and

Embeddings for Aspect Based Sentiment Analysis", SemEval 2016, pp. 312–317.

[17] T.T. Vu, H.T. Pham, C.T. Luu, and Q.T. Ha, "A feature-based opinion mining model on product reviews in Vietnamese", Studies in Computational Intelligence, Vol. 381, 2011, pp. 23–33.

[18] H.S. Le, T.L. Van, and T.V. Pham, "Aspect analysis for opinion mining of Vietnamese text", ACOMP 2015, pp. 118–123.

[19] N.T.M. Huyen, N.V. Hung, N.T. Quyen, V.X. Luong, T.M. Vu, N.X. Bach, and L.A. Cuong, "VLSP Shared Task: Sentiment Analysis. Journal of Computer Science and Cybernetics", Vol. 34, 2018, No. 4, pp. 295–310.

[20] D.V. Thin, N.D. Vu, N.V. Kiet, and N.L.T. Ngan, "A transformation method for aspect-based sentiment analysis", Journal of Computer Science and Cybernetics, Vol. 34, 2018, No. 4, pp. 323–333.

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

[22] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for Russian language," arXiv [cs.CL], 2019.

[23] A. Sarkar, S. Reddy, and R. S. Iyengar, "Zero-Shot Multilingual Sentiment Analysis using Hierarchical Attentive Network and BERT," presented at the NLPIR 2019: 2019 the 3rd International Conference on Natural Language Processing and Information Retrieval, Jun. 2019, doi: 10.1145/3342827.3342850.

[24] I. F. Putra and A. Purwarianti, "Improving Indonesian text classification using multilingual language model," arXiv [cs.CL], 2020.

[25] S. Rönnqvist, J. Kanerva, T. Salakoski, and F. Ginter, "Is multilingual BERT fluent in language generation?," arXiv [cs.CL], 2019.

[26] Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog," presented at the Proceedings of the 2019 Conference of the North, 2019, doi: 10.18653/v1/n19-1380.

[27] N.T.T. Thuy, N.X. Bach, and T.M. Phuong, "Cross-Language Aspect Extraction for Opinion Mining", KSE 2018, pp. 67–72.

[28] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 1907. Roberta: A robustly optimized bert pretraining approach. arXiv 2019. arXiv preprint arXiv:1907.11692.

[29] Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in véc-tơ space. arXiv preprint arXiv:1301.3781.

[30] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. and Klingner, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

[31] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

[32] Alex Graves. 2013. Generating sequences with recurrent neural networks. CoRR, abs/1308.0850.

[33] Hasim Sak, AndrewW. Senior, and Franoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In INTERSPEECH, pages 338–342.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

[35] Alexis Conneau, Douwe Kiela, Holger Schwenk, Lo¨ıc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

[36] Sun, C., Huang, L. and Qiu, X., 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588.

USING BERT AND AUXILIARY SENTENCE FOR ASPECT CATEGORY EXTRACTION IN VIETNAMESE TEXT

Abstract: Aspect extraction is one core task of aspect-based opinion mining, to identify and classify opinion words or phrases for each feature from customer reviews. Most of the previous studies on aspect extraction and aspect-based opinion mining are for English texts, and there are very few studies for Vietnamese. Studies for Vietnamese with higher accuracy are often based on supervised learning methods or based on deep learning, with models dependent on context-independent word embedding (such as word2vec). This paper presents an aspect extraction method based on modeling capabilities with contextualized word embeddings, using pre-trained language models such as BERT. Unlike previous studies that used a single input review sentence to extract aspect category, our proposed method uses auxiliary sentences constructed from aspect keywords to take advantage of known key information, combined with the original input sentence to create a pair of input sentences for BERT, then do classification. The proposed BERT-based model has an extra linear layer for classification, tuned along with the auxiliary sentence, showing good results on the annotated corpus of aspect categories which are collected from restaurant reviews/comments on social networks in the Vietnamese language.

Keywords: aspect category extraction, opinion mining for Vietnamese, pre-trained language model, BERT.

Nguyễn Ngọc Điệp. Nhận học vị Tiến sĩ năm 2017. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, an toàn thông tin, xử lý ngôn ngữ tự nhiên.

Nguyễn Thị Thanh Thủy. Nhận học vị Thạc sĩ năm 2009. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, xử lý ngôn ngữ tự nhiên.