TRÍCH XUẤT DANH MỤC KHÍA CẠNH SỬ DỤNG BERT VỚI HÀM MẤT MÁT CÂN BẰNG

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

Tóm tắt: Trích xuất danh mục khía cạnh (*aspect category extraction*) là nhiệm vụ đầu tiên trong bài toán khai thác quan điểm dựa trên khía cạnh (*aspect-based opinion mining*). Đây là một nhiệm vụ khó khăn vì người dùng thường sử dụng các từ khóa khác nhau để diễn tả về cùng một khía cạnh hoặc nhiều khi chỉ dùng các từ ngụ ý đề cập đến khía cạnh. Các phương pháp học máy có giám sát nói chung được đánh giá là có độ chính xác cao, tuy nhiên thường tốn kém nhiều công sức trong việc gán nhãn dữ liệu huấn luyện, đặc biệt là cho các miền lĩnh vực mới. Hơn nữa, các phương pháp này thường yêu cầu phải có kiến thức chuyên gia giúp trích chọn ra được các đặc trưng thủ công hữu ích đối với miền lĩnh vực nghiên cứu. Bài báo này trình bày đề xuất một phương pháp cải tiến sử dụng mô hình học sâu dựa trên BERT để giải quyết và nâng cao hiệu năng cho nhiệm vụ trích xuất danh mục khía cạnh. Mô hình đề xuất tự học các đặc trưng từ chuỗi dữ liệu văn bản đầu vào và biểu diễn hiệu quả nhờ BERT. Ngoài ra, để khắc phục vấn đề mất cân bằng dữ liệu giữa các nhãn lớp, chúng tôi đề xuất sử dụng các hàm mất mát cân bằng (*balanced loss functions*). Kết quả thực nghiệm cho thấy mô hình đề xuất có hiệu năng vượt trội hơn, với trung bình độ đo F_1 cao nhất đạt 77%.

Từ khóa: trích xuất danh mục khía cạnh, học máy, học sâu, BERT, hàm mất mát cân bằng.

I. GIỚI THIỆU

Trong những năm gần đây, khai thác quan điểm dựa trên khía cạnh (*aspect-based opinion mining*) là một chủ đề nhận được rất nhiều quan tâm từ cộng đồng nghiên cứu xử lý ngôn ngữ tự nhiên (*natural language processing*) và khai phá dữ liệu (*data mining*). Không giống như phân loại cảm xúc *(sentiment classification)*, trong đó xác định cảm xúc chung cho một văn bản có thể hiện quan điểm/ý kiến, khai thác 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. Cụ thể, khai thác quan điểm dựa trên khía cạnh bao gồm hai nhiệm vụ chính là: 1) Trích xuất danh mục khía cạnh, trong đó xác định các loại danh mục khía cạnh (các cặp thực thể và thuộc tính) có diễn tả ý kiến/quan điểm trong văn bản; và 2) Phân loại phân cực cảm xúc, trong đó gán một nhãn phân cực (tích cực, tiêu cực hoặc trung tính) cho mỗi loại khía cạnh đã được xác định. Ví dụ, cho một câu văn bản đầu vào là: "*Tôi thấy đồ ăn ở đây khá ngon, nhưng lại hơi xa khu trung tâm*", thì đầu ra của một hệ thống khai thác quan điểm dựa trên khía cạnh sẽ là: 1) Có hai loại danh mục khía cạnh người dùng nhắc đến: đồ ăn và vị trí của nhà hàng. 2) Hai nhãn phân loại phân cực cảm xúc tương ứng với từng danh mục khía cạnh: tích cực (hài lòng) với "đồ ăn", và tiêu cực (không hài lòng) với "vị trí" của nhà hàng.

Thực tế hiện nay, nguồn dữ liệu web được phát triển vô cùng phong phú và đa dạng, trong đó ngày càng có nhiều hơn những bình luận/đánh giá của người dùng về các sản phẩm/dịch vụ mà họ đã từng mua/sử dụng với mức độ chi tiết đến từng khía cạnh/đặc trưng của sản phẩm/dịch vụ. Việc phân tích quan điểm của người dùng đối với các sản phẩm/dịch vụ theo khía cạnh/đặc trưng đóng vai trò quan trọng cả với người dùng là khách hàng, người bán hàng và nhà sản xuất. Kết quả phân tích sẽ giúp khách hàng lựa chọn được sản phẩm/dịch vụ tốt; giúp người bán hàng và nhà sản xuất nắm được thị hiếu của khách hàng, xu hướng thị trường; cũng từ đó, giúp nhà sản xuất định hướng thiết kế, phát triển các dòng sản phẩm/dịch vụ tiếp theo.

Có thể nhận thấy, nhiệm vụ trích xuất danh mục khía cạnh đóng rất vai trò quan trọng trong khai thác quan điểm dựa trên khía cạnh, bởi hai lý do sau. (1) Khi trích xuất được chính xác khía cạnh người dùng muốn nói đến trong văn bản, thì mới có thể biết được ý kiến/quan điểm của họ về thuộc tính cụ thể nào của sản phẩm/dịch vụ được đề cập đến, thay vì chỉ biết được ý kiến/quan điểm về sản phẩm/dịch vụ nói chung. Và (2) độ chính xác của phân loại cảm xúc phụ thuộc vào độ chính xác của việc trích xuất danh mục khía cạnh trong khai thác quan điểm dựa trên khía cạnh.

Trong một nghiên cứu trước của nhóm [4], chúng tôi đã giải quyết nhiệm vụ trích xuất danh mục khía cạnh sử dụng các phương pháp học máy có giám sát truyền thống, với đề xuất sử dụng thêm tài nguyên sẵn có từ các ngôn ngữ giàu tài nguyên (như tiếng Anh) cho các ngôn ngữ nghèo tài nguyên (như tiếng Việt). Một yêu cầu quan trọng khi sử dụng các phương pháp học máy truyền thống là cần phải có kiến thức chuyên gia giúp trích chọn ra được các đặc trưng thủ công hữu ích đối với miền lĩnh vực đang nghiên cứu. Nghiên cứu ở đây khác nghiên cứu trước, đó là chúng tôi sử dụng mô hình học sâu dựa trên BERT để giải quyết và

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

cải tiến hiệu năng cho nhiệm vụ trích xuất danh mục khía cạnh. 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. Ngoài ra, để khắc phục vấn đề mất cân bằng dữ liệu giữa các nhãn lớp, chúng tôi đề xuất sử dụng các hàm mất mát cân bằng. Kết quả thực nghiệm cho thấy mô hình học sâu đề xuất sử dụng BERT và các hàm mất mát cân bằng có hiệu năng vượt trội so với mô hình học máy trước [4].

Đóng góp của nghiên cứu này gồm hai phần. Thứ nhất, chúng tôi đề xuất mô hình học sâu hiệu quả dựa trên kiến trúc BERT và các hàm mất mát cân bằng, với khả năng học đặc trưng tốt cho chuỗi dữ liệu văn bản đầu vào và hoạt động tốt trên tập dữ liệu mất cân bằng về danh mục khía cạnh. Thứ hai, chúng tôi trình bày tính hiệu quả của phương pháp đề xuất bằng việc thực hiện một chuỗi các thực nghiệm trên một tập dữ liệu kết hợp từ tập dữ liệu tự gán nhãn cho ngôn ngữ tiếng Việt cùng với tập dữ liệu được gán nhãn sẵn từ ngôn ngữ tiếng Anh. Kết quả cho thấy phương pháp đề xuất mới cùng với việc sử dụng các biến thể của hàm mất mát cân bằng đạt trung bình độ đo F_1 cao nhất là 77%, tăng 5% so với nghiên cứu trước [4].

Phần còn lại của bài báo được tổ chức như sau. Phần II mô tả các nghiên cứu liên quan. Phần III trình bày đề xuất phương pháp thực hiện trích xuất danh mục khía cạnh. Thông tin về bộ dữ liệu và các kết quả thực nghiệm được trình bày trong phần Phần IV và Phần V. Cuối cùng, Phần VI là kết luận bài báo và định hướng nghiên cứu.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Phần này trình bày các nghiên cứu liên quan đến các phương pháp trích xuất danh mục khía cạnh và các nghiên cứu liên quan về khai thác quan điểm dựa trên khía cạnh trong tiếng Việt.

A. Trích xuất danh mục khía cạnh

Các nghiên cứu về trích xuất danh mục khía cạnh có thể được chia thành hai loại chính: các phương pháp dựa trên phân cụm và các phương pháp dựa trên phân loại.

Các phương pháp dựa trên phân cụm thường được sử dụng trong các trường hợp không thể xác định được chính xác danh sách các loại khía cạnh và/hoặc không có dữ liệu được chú thích thông tin về các loại danh mục khía cạnh. Ví dụ, trong nghiên cứu [1], tác giả He và cộng sự thực hiện trích xuất tất cả các thuật ngữ chỉ khía cạnh từ một kho văn bản các bài đánh giá/bình luận, (ví dụ: "thịt bò", "thịt lợn",…). Sau đó, các cụm từ có ý nghĩa tương tự nhau sẽ được nhóm chung vào thành một loại danh mục khía cạnh, (ví dụ: nhóm "nấm", "thịt bò", "thịt lợn", và "cà chua" thành một loại danh mục khía cạnh là đồ ăn). Các phương pháp theo cách tiếp cận này thường sử dụng các thuật toán học không giám sát hoặc bán giám sát với kỹ thuật bootstrapping [1], [2], [3]. Các phương pháp khác được sử dụng phổ biến cho trích xuất cụm từ khía cạnh là phương pháp dựa trên luật và mô hình chủ đề. Trong khi phương pháp dựa trên luật sử dụng tính phụ thuộc về cú pháp hoặc các mối quan hệ từ (ví dụ, mối quan hệ giữa từ thể hiện ý kiến và từ khía cạnh) để trích xuất các thuật ngữ khía cạnh [5], thì phương pháp mô hình chủ đề coi các loại danh mục khía cạnh như là các chủ đề, được phân phối qua các từ (khía cạnh) trong kho văn bản [6], [7], [8].

Các phương pháp dựa trên phân loại thường được được sử dụng trong các trường hợp biết được chính xác danh sách các loại khía cạnh và có dữ liệu đã được chú thích. Các phương pháp này xác định các loại khía cạnh của các bài đánh giá mà không cần trích xuất các thuật ngữ khía cạnh. Phương pháp thực hiện mô hình hóa nhiệm vụ trích xuất như là một bài toán phân loại đa nhãn (multi-label classification) hoặc nhiều bài toán phân loại nhị phân trong đó mỗi nhãn tương ứng với một loại khía cạnh [9], [10], [11].

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 và phân loại đa nhãn cùng với việc sử dụng các hàm mất mát cân bằng để trích xuất danh mục khía cạnh. Mô hình phân loại đa nhãn sẽ xác định xem văn bản đầu vào đề cập đến các loại danh mục khía cạnh nào.

B. Khai thác quan điểm dựa trên khía cạnh tiếng Việt

Phần lớn các nghiên cứu trước đây về phân tích cảm xúc và khai phá quan điểm cho tiếng Việt đều tập trung vào phân loại cảm xúc. Tác giả Duyên và cộng sự [12] mô tả một chuỗi các thực nghiệm về phân loại cảm xúc dựa trên huấn luyện cho tiếng Việt, tập trung vào một số phương pháp trích xuất đặc trưng và các thuật toán học có giám sát, như Naive Bayes, mô hình Entropy cực đại và máy véc-tơ hỗ trợ. Tác giả Hà và cộng sự [13] mô tả một phương pháp sử dụng các đặc trưng bag-of-bigram trong lifelong learning framework cho phân loại cảm xúc chéo lĩnh vực (cross-domain) cho tiếng Việt. Tác giả Bách và Phương [14] trình bày một phương pháp học có giám sát yếu phân loại cảm xúc cho ngôn ngữ nghèo tài nguyên. Phương pháp này khai thác thông tin xếp hạng tổng thể của bài đánh giá như là thông tin bổ sung để huấn luyện bộ phân loại cảm xúc bán giám sát và được chứng minh là có hiệu quả trên hai bộ dữ liệu tiếng Nhật và tiếng Việt. Các tác giả Kiều và Phạm [15] giới thiệu một hệ thống dựa trên luật cho phân loại cảm xúc tiếng Việt bằng cách sử dụng Gate framework. Tác giả Phú và cộng sự [16] đề xuất mô hình valence-totaling cho phân loại cảm xúc tiếng Việt. Phương pháp của họ đạt được độ chính xác là 63,9% trên một kho văn bản tiếng Việt gồm 15.000 tài liệu có ý kiến tích cực và 15.000 tài liệu có ý kiến tiêu cực.

Có một số ít các nghiên cứu về khai phá quan điểm dựa trên khía cạnh cho tiếng Việt. Tác giả Lê và cộng sự [17] trình bày một phương pháp học bán giám sát cho trích xuất và phân loại các thuật ngữ khía cạnh trong văn bản tiếng Việt. Đầu tiên, phương pháp trích xuất tất cả các từ đã được tách từ (các từ ghép có ý nghĩa) từ một kho các bài đánh giá và chọn các từ có tần số xuất hiện nhiều nhất, tiếp theo chúng sẽ được gán nhãn thủ công và được dùng như là "hạt giống" của thuật toán. Sau đó, cây quyết định sẽ được xây dựng để phân loại các thuật ngữ khía cạnh. Tác giả Vũ và cộng sự [18] mô tả một nghiên cứu về khai phá quan điểm dựa trên khía cạnh trên các bài đánh giá sản phẩm bằng tiếng Việt. Các từ thể hiện khía cạnh (rõ ràng hoặc tiềm ẩn) và các từ thể hiện ý kiến/quan điểm được trích xuất bằng

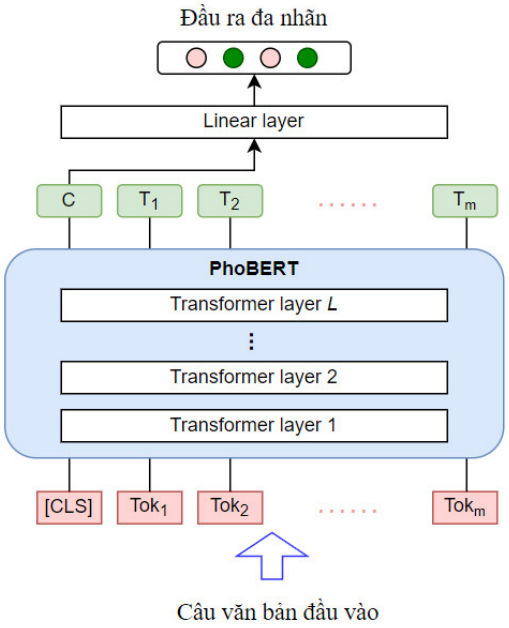
các luật theo cú pháp tiếng Việt. Gần đây, một số nghiên cứu đã được trình bày bằng thuật toán học có giám sát [19, 20] trong một nhiệm vụ chung về phân tích cảm xúc dựa trên khía cạnh tiếng Việt tại hội thảo quốc tế lần thứ năm về Xử lý ngôn ngữ và tiếng Việt (VLSP 2018) [19].

Nghiên cứu của chúng tôi khác với những nghiên cứu khác ở chỗ chúng tôi đề xuất một phương pháp mới dựa trên học sâu và sử dụng các hàm mất mát cân bằng để giải quyết nhiệm vụ. Hơn nữa, chúng tôi xác định các loại danh mục khía cạnh trong một câu thay vì toàn bộ bài đánh giá. Chúng tôi tin rằng nhiệm vụ ở cấp độ câu là thực tế hơn và có thể áp dụng trong các ứng dụng thế giới thực.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Phần này trình bày đề xuất phương pháp trích xuất danh mục khía cạnh sử dụng mô hình học sâu dựa trên kiến trúc BERT. Đối với mỗi câu có diễn tả quan điểm, nhiệm vụ là cần xác định xem câu đó nói đến loại danh mục khía cạnh nào. Cụm từ chỉ đến loại danh mục khía cạnh có thể được nói tường minh trong câu hoặc có thể chỉ ở dạng ngầm định.

Giả sử có một tập văn bản thô ***D***. Cho *s* là một câu trong tập văn bản ***D***. Thực hiện phân đoạn câu *s* thành chuỗi có *m* từ (token) sử dụng nhúng từ, cụ thể là byte pair encoding (gồm cả 2 token đặc biệt để đánh dấu vị trí bắt đầu [CLS] và kết thúc [SEP] câu). Giả sử có tập *K* danh mục khía cạnh. Mỗi khía cạnh được thể hiện tương ứng bằng 1 one-hot véc-tơ độ dài *K*. Đầu ra của mô hình là một véc-tơ độ dài *K* có giá trị thuộc {0, 1}. Do mỗi câu có thể chứa một hoặc nhiều danh mục khía cạnh nên số lượng các phần tử có giá trị bằng 1 trong véc-tơ đầu ra có thể là 1 hoặc nhiều hơn 1. Ví dụ: Cho một câu đầu vào *s* là "*Tôi thấy đồ ăn ở đây khá ngon, nhưng lại hơi xa khu trung tâm*", thì véc-tơ đầu ra tương ứng của câu là {0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1} (với giả thiết là có 12 danh mục khía cạnh nói về nhà hàng).



Hình 1. Kiến trúc tổng thể của mô hình trích xuất danh mục khía cạnh

Sơ đồ tổng quan về mô hình đề xuất được trình bày trong Hình 1. Mô hình có 2 thành phần là BERT và mô đun phân lớp. Chuỗi văn bản đầu vào được chuyển thành chuỗi *m* token sử dụng lớp embeddings có sẵn trong mô hình pre-trained BERT. Đầu ra của BERT là chuỗi embeddings độ dài *d* (*d* là số unit ẩn có trong mô hình BERT), là véc-tơ trạng thái ẩn tại lớp cuối cùng trong BERT. Tiếp sau là 1 lớp dropout để làm giảm vấn đề quá khớp dữ liệu, và 1 lớp kết nối đầy đủ (*Fully Connected Layer*) có số unit đầu ra là *K* (độ lớn của danh mục khía cạnh) với hàm kích hoạt là sigmoid.

A. *BERT*

BERT (*Bidirectional Encoder Representations from Transformers*) là một kỹ thuật học máy dựa trên các Transformers, được phát triển vào năm 2018 [22]. BERT là một mô hình học sẵn (pre-trained model), học được các véc-tơ đại diện theo ngữ cảnh 2 chiều của từ, nghĩa là BERT tạo ra các biểu diễn từ theo ngữ cảnh dựa trên các từ trước và sau đó trong câu để dẫn đến một mô hình ngôn ngữ với ngữ nghĩa phong phú hơn so với các phương pháp biểu diễn khác trước đây (Word2vec, FasText, hay Glove). Ngay sau khi công bố, BERT được coi là bước đột phá trong công nghệ xử lý ngôn ngữ tự nhiên.

Kiến trúc của mô hình BERT là dạng kiến trúc đa tầng, bao gồm nhiều lớp (blocks) Bidirectional Transformer encoder. BERT có 2 phiên bản kiến trúc là BERT-*base* có 12 lớp transformer và BERT-*large* có 24 lớp transformer. Giả sử gọi *L* là số lớp transformer được sử dụng, *H* là số các lớp ẩn, *A* là số head ở lớp attention, thì kích thước của 2 mô hình tương ứng với 2 phiên bản kiến trúc là:

- BERT-*base*: *L*=12, *H*=768, *A*=12, Total Parameters=110M
- BERT-*large*: *L*=24, *H*=1024, *A*=16, Total Parameters=340M

PhoBERT [23] là một mô hình ngôn ngữ được huấn luyện sẵn riêng cho tiếng Việt, với khoảng 20G dữ liệu huấn luyện từ Wikipedia và kho các trang tin tức tiếng Việt. Kiến trúc của mô hình PhoBERT tương tự với mô hình BERT, là bộ mã hóa transformer hai chiều *L* tầng (*L-layer bidirectional Transformer encoder*) [24]. Trong nghiên cứu này, chúng tôi sử dụng PhoBERT để mã hóa các thông tin ngữ cảnh cho bài toán trích xuất danh mục khía cạnh. Các véc-tơ ẩn của tầng cuối từ PhoBERT được sử dụng làm biểu diễn chung của mỗi từ (token) trong câu đầu vào *s*.

B. *Biểu diễn đầu vào*

Đầu vào văn bản cho mô hình BERT [22] là một câu hoặc một chuỗi văn bản các token. Như vậy, mỗi câu (hoặc chuỗi văn bản) sẽ bao gồm 1 tập các token, mỗi token sẽ đại diện cho 1 từ. Có hai token đặc biệt được thêm vào tập các token: token phân loại [CLS], được thêm vào đầu câu (hoặc chuỗi), và token phân tách [SEP], để đánh dấu phần kết thúc câu (hoặc chuỗi). Tập các token này sau đó được xử lý thông qua ba lớp nhúng khác nhau có cùng kích thước, cuối cùng được tổng hợp lại với nhau và chuyển đến lớp mã hóa, bao gồm lớp nhúng từ (token embeddings),

lớp nhúng phân đoạn (segmentation embeddings) và lớp nhúng vị trí (position embeddings). Vị trí ở đây là vị trí tương ứng của các token trong câu (hoặc chuỗi văn bản).

C. *Mất cân bằng lớp và các hàm mất mát cân bằng*

Trong các bài toán phân loại, mất cân bằng dữ liệu giữa các nhãn lớp là một trong những nguyên nhân dẫn đến giảm hiệu quả việc phân loại, đặc biệt với các bài toán phân loại đa nhãn. Có hai phương pháp để giải quyết sự mất cân bằng này, một là lấy lại mẫu (hoặc có thể bổ sung thêm các mẫu có số lượng còn thấp), hai là đặt lại trọng số cho tập dữ liệu huấn luyện. Tuy nhiên, trong bài toán đa nhãn, các phương pháp này cũng không thực sự hiệu quả. Ví dụ, trong trường hợp ngoài sự mất cân bằng lớp còn có sự phụ thuộc giữa các nhãn, nếu lấy thêm dữ liệu bổ sung sẽ dẫn đến hiện tượng quá nhiều dữ liệu với các nhãn phổ biến.

Trong nghiên cứu về bài toán nhận dạng đối tượng ảnh [25], một phương pháp để giải quyết vấn đề mất cân bằng dữ liệu là sử dụng các hàm mất mát cân bằng, trong đó có hàm mất mát tiêu điểm (focal loss) đã được sử dụng và khắc phục được vấn đề mất cân bằng giữa các nhóm foreground và background của các đối tượng ảnh. Cách thực hiện là xác định lại hàm lỗi entropy chéo tiêu chuẩn để giảm trọng số lỗi được gán cho các mẫu đã được phân loại tốt (dễ dự đoán). Trong nghiên cứu này, chúng tôi áp dụng các hàm focal loss khi phân loại văn bản đa nhãn và thử nghiệm với một số biến thể của nó. Kết quả cho thấy hàm focal loss được cân bằng về phân phối giúp giải quyết được cả vấn đề mất cân bằng lớp và liên kết (phụ thuộc) nhãn, hoạt động tốt hơn các hàm mất cân bằng thường được sử dụng.

Phần này giới thiệu tóm tắt về một số hàm mất mát thường được sử dụng trong bài toán phân loại văn bản đa nhãn.

Binary Cross Entropy. Hàm mất mát entropy chéo nhị phân (Binary Cross Entropy - BCE) thường được sử dụng trong phân loại văn bản đa nhãn trong xử lý ngôn ngữ tự nhiên [26]. Giả sử có tập dữ liệu gồm N mẫu huấn luyện $\{(x^1,y^1), (x^2,y^2), ..., (x^N,y^N)\}$, mỗi mẫu có tương ứng một véc-tơ đa nhãn thực $y^k=[y_1{}^k,...,y_C{}^k]\in\{0,1\}^C$ (với C là số lượng các lớp) và một véc-tơ đa nhãn dự đoán (là nhãn đầu ra của bộ phân lớp) $z^k=[z_1{}^k,...,z_C{}^k]\in R$. Hàm mất mát BCE được xác định bằng công thức sau:

$$L_{BCE}=\begin{cases}-\log\left(p_i^k\right) & nếu\ y_i^k=1\\ -\log\left(1-p_i^k\right) & ngược\ lại\end{cases} \quad (1)$$

Hàm sigmoid được sử dụng để tính $p_i{}^k$, với $p_i{}^k=\sigma(z_i{}^k)$.

Focal loss. Focal loss (FL) là hàm mất mát được giới thiệu trong nghiên cứu [25], và về sau đã được áp dụng hiệu quả trong các bài toán có sự mất cân bằng dữ liệu lớn giữa các lớp (ví dụ số lượng các nhãn âm lớn hơn rất nhiều các nhãn dương). Focal loss được tính toán bằng cách nhân một hệ số điều biến với hàm mất mát BCE. Hệ số điều biến có tác dụng rất lớn trong việc điều chỉnh ảnh hưởng của nhãn lên đồng thời hàm mất mát và gradient descent. Do đó, việc sử dụng focal loss sẽ làm giảm mức độ tập trung trong trường hợp dễ dự đoán (có nghĩa là dữ liệu đã được học tốt rồi) và sẽ tăng mức độ tập trung vào trường hợp khó dự đoán (nghĩa là dữ liệu khó học hơn). Đối với bài toán phân loại đa nhãn, hàm focal loss được xác định như sau:

$$L_{FL}=\begin{cases}-\left(1-p_i^k\right)^\gamma\log\left(p_i^k\right) & nếu\ y_i^k=1\\ -\left(p_i^k\right)^\gamma\log\left(1-p_i^k\right) & ngược\ lại\end{cases} \quad (2)$$

Class-balanced focal loss. Trong nghiên cứu [27], hàm focal loss cân bằng theo lớp (Class-balanced focal loss - CB) được xây dựng bằng cách ước tính số lượng mẫu hiệu quả. Ý tưởng ở đây là tiếp tục cân bằng lại trọng số trong hàm FL để giúp giảm thông tin dư thừa của các lớp nhiều mẫu. Đối với các bài toán phân loại đa nhãn, mỗi nhãn có tần suất tổng thể n_i có hệ số cân bằng riêng:

$$r_{CB}=\frac{1-\beta}{1-\beta^{n_i}} \quad (3)$$

trong đó, $\beta\in[0, 1)$ kiểm soát tốc độ tăng hiệu quả. Hàm mất mát sẽ được tính như sau:

$$L_{CB}=\begin{cases}-r_{CB}\left(1-p_i^k\right)^\gamma\log\left(p_i^k\right) & nếu\ y_i^k=1\\ -r_{CB}\left(p_i^k\right)^\gamma\log\left(1-p_i^k\right) & ngược\ lại\end{cases} \quad (4)$$

Distribution-balanced loss. Trong nghiên cứu [28], hàm mất mát cân bằng phân phối (Distribution-balanced loss - DB) được đề xuất nhằm làm giảm thông tin dư thừa trong trường hợp đồng xuất hiện nhãn (rất quan trọng trong kịch bản nhiều nhãn), sau đó xác định rõ trọng số thấp hơn cho các trường hợp nhãn âm dễ dự đoán. Hàm DB được xác định bằng cách tích hợp trọng số cân bằng lại và điều hòa dung sai tiêu cực (negative tolerant regularization - NTR), như sau:

$$L_{DB}=\begin{cases}-\hat{r}_{DB}\left(1-q_i^k\right)^\gamma\log\left(q_i^k\right) & nếu\ y_i^k=1\\ -\hat{r}_{DB}\dfrac{1}{\lambda}\left(q_i^k\right)^\gamma\log\left(1-q_i^k\right) & ngược\ lại\end{cases} \quad (5)$$

trong đó, $q_i{}^k=\sigma(z_i{}^k-v_i)$ với các mẫu dương, và $q_i{}^k=\sigma(\lambda(z_i{}^k-v_i))$ với các mẫu âm. λ là hệ số tỷ lệ và v_i là độ lệch lớp cụ thể được đưa ra để giảm ngưỡng cho phần cuối của các lớp tránh bị triệt tiêu quá mức.

D. *Huấn luyện mô hình*

Mô hình được huấn luyện lấy dữ liệu câu đầu vào từ tập dữ liệu huấn luyện. Danh mục khía cạnh tương ứng với câu đầu vào được chuyển thành véc-tơ nhị phân có độ dài K bằng với số lượng danh mục khía cạnh, với mỗi phần tử thuộc {0, 1}. Với bài toán phân lớp đa nhãn, đầu ra là véc-tơ nhị phân, hàm mất mát được sử dụng cho thử nghiệm đầu tiên (baseline) là hàm mất mát entropy chéo nhị phân (Binary Cross Entropy). Hàm tối ưu sử dụng khi huấn luyện mô hình là Adam.

Để khắc phục vấn đề mất cân bằng giữa các lớp, chúng tôi sử dụng các hàm mất mát cân bằng phân phối dựa trên các biến thể của hàm focal loss. Cách thức là định hình lại hàm lỗi entropy chéo tiêu chuẩn để làm giảm trọng số lỗi được gán cho các mẫu được phân loại tốt (dễ dự đoán).

Mô hình được tùy chỉnh trên tập dữ liệu kiểm tra, với các tham số learning rate, batch size và dropout được tối ưu trong khoảng tương ứng: {1e-5, 2e-5, 3e-5, 4e-5, 5e-5}, {8,

16, 32, 64}, {0.1, 0.2, 0.3, 0.4, 0.5}. Giá trị tốt nhất của các tham số được chọn cho mô hình tối ưu tương ứng là: learning rate = 2e-5, batch size = 32, dropout = 0.3.

IV. TẬP DỮ LIỆU

Phần này giới thiệu về các bộ dữ liệu được sử dụng trong thực nghiệm của nghiên cứu, bao gồm tập dữ liệu tiếng Việt chúng tôi tự xây dựng trong nghiên cứu trước của nhóm về trích xuất danh mục khía cạnh [4] và tập dữ liệu tiếng Anh từ SemEval-2016 Task 5 [10]. Phần sau sẽ trình bày tóm tắt lại quá trình xây dựng cũng như các thống kê cụ thể về tập dữ liệu.

1) Tập dữ liệu tiếng Việt [4]*:* Tập dữ liệu tiếng Việt được thu thập từ Foody (*https://www.foody.vn/*). Đây là một trang web lớn và phổ biến nhất Việt Nam, nơi người dùng có thể tìm kiếm, đánh giá/bình luận và đặt hàng, không chỉ riêng với đồ ăn uống mà còn có cả dịch vụ du lịch, làm đẹp và các dịch vụ khác. Chúng tôi thực hiện trích xuất các bài đánh giá từ một số nhà hàng ở Việt Nam (hầu hết ở Hà Nội và thành phố Hồ Chí Minh). Sau đó tiến hành một số bước tiền xử lý trên dữ liệu thô, bao gồm làm sạch dữ liệu, phát hiện ranh giới câu. Kết quả thu được tập dữ liệu bao gồm 575 bài đánh giá với 3796 câu tiếng Việt.

Có hai người chú thích thực hiện gán nhãn các danh mục khía cạnh cho từng câu sau khi đã được tiền xử lý. Nếu hai người chú thích không đồng ý kiến về một nhãn được gán, thì sẽ có một người thứ ba kiểm tra và đưa ra quyết định cuối cùng. Những người chú thích là sinh viên chuyên ngành Công nghệ thông tin của Học viện Công nghệ bưu chính viễn thông (hai sinh viên đại học và một sau đại học) có kiến thức nền tảng cơ bản về xử lý ngôn ngữ tự nhiên và học máy. Hệ số Kappa của Cohen được sử dụng để đo mức độ tương đồng ý kiến giữa các chú thích:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (6)$$

trong đó $Pr(a)$ là độ tương đồng ý kiến giữa hai người chú thích, và $Pr(e)$ là xác suất giả thiết có ý kiến khác nhau. Do mỗi câu có thể được gán với nhiều nhãn danh mục khía cạnh, nên chúng tôi tính hệ số Kappa cho từng loại danh mục, và lấy kết quả tính trung bình. Hệ số Kappa của tập dữ liệu tiếng Việt chúng tôi thực hiện gán nhãn là 0,83, là một giá trị được coi là có độ tương đồng ý kiến rất tốt [21].

Tương tự như SemEval-2016 Task 5 [10], chúng tôi xem xét 12 loại khía cạnh được đại diện bởi 12 bộ (thực thể, thuộc tính). Hình 2 trình bày ví dụ về một bài đánh giá có chú thích gồm ba câu trong tập văn bản được xây dựng. Câu đầu tiên bình luận về chất lượng và giá của đồ ăn (danh mục FOOD#QUALITY và FOOD#PRICES). Câu thứ hai bình luận về thái độ phục vụ của nhân viên (danh mục SERVICE#GENERAL). Câu cuối cùng nhận xét về không gian của nhà hàng (danh mục AMBIENCE#GENERAL).

```
<Review rid="bc8">
    <Sentence id="bc8:0">
        <Text> Đồ_ăn ngon nhưng khá đắt.
        </Text>
        <Categories> FOOD#QUALITY; FOOD#PRICES
        </Categories>
    </Sentence>
    <Sentence id="bc8:1">
        <Text> Nhân_viên lịch_sự, phục_vụ nhanh.
        </Text>
        <Categories> SERVICE#GENERAL
        </Categories>
    </Sentence>
    <Sentence id="bc8:2">
        <Text> Tuy_nhiên, không_gian ở đây hơi chật_chội.
        </Text>
        <Categories> AMBIENCE#GENERAL
        </Categories>
    </Sentence>
</Review>
```

Hình 2. Các câu trong một bài đánh giá được chú thích trong tập dữ liệu tiếng Việt

2) Tập dữ liệu tiếng Anh: Chúng tôi sử dụng tập dữ liệu tiếng Anh (cả dữ liệu huấn luyện và dữ liệu kiểm tra) từ SemEval-2016 Task 5 [10] làm nguồn dữ liệu bổ sung cho phương pháp đề xuất. Như trình bày trong Bảng I, tập dữ liệu tiếng Anh bao gồm 440 bài đánh giá với 2676 câu. Như vậy tổng cộng có 1015 bài đánh giá với 6472 câu cho cả bộ dữ liệu tiếng Việt và tiếng Anh.

Bảng I. Thông tin thống kê trong hai tập dữ liệu

	Tiếng Việt	Tiếng Anh	Tổng số
Số bài đánh giá	575	440	1015
Số câu	3796	2676	6472

Bảng II. Danh mục khía cạnh và tần số xuất hiện của chúng trong hai tập dữ liệu

STT	Danh mục khía cạnh	Tiếng Việt	Tiếng Anh
1	RESTAURANT#GENERAL	233	564
2	RESTAURANT#PRICES	132	101
3	RESTAURANT#MISCELLANEOUS	194	131
4	FOOD#QUALITY	1357	1162
5	FOOD#STYLE_OPTIONS	586	192
6	FOOD#PRICES	207	113
7	DRINKS#QUALITY	307	69
8	DRINKS#STYLE_OPTIONS	75	44
9	DRINKS#PRICES	56	24
10	SERVICE#GENERAL	487	604
11	AMBIENCE#GENERAL	516	321
12	LOCATION#GENERAL	215	41
	Tổng	**4365**	**3366**

Bảng II trình bày chi tiết 12 loại danh mục khía cạnh và tần số xuất hiện của chúng trong bộ dữ liệu tiếng Việt và tiếng Anh. Trong cả 2 tập dữ liệu, 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.

Trong quá trình thực hiện thực nghiệm, tất cả các câu trong tập dữ liệu tiếng Anh được dịch sang tiếng Việt bằng công cụ Google Translate *(https://translate.google.com/)*. Việc dịch trực tiếp (qua phiên dịch viên) trong thực tế sẽ mang lại kết quả dịch chính xác hơn so với dịch máy tự động, tuy nhiên việc này sẽ làm tốn kém rất nhiều thời gian với lượng dữ liệu lớn, đồng thời cũng khó khăn khi chuyển đổi giữa các ngôn ngữ khác nhau. Do vậy, trong nghiên cứu này chúng tôi chọn phương pháp dịch máy.

V. CÁC THỰC NGHIỆM VÀ KẾT QUẢ

Phần này sẽ trình bày thiết lập thực nghiệm, các kết quả thực nghiệm và phân tích kết quả.

A. *Thiết lập thực nghiệm*

Chúng tôi chia ngẫu nhiên tập dữ liệu tiếng Việt thành 10 phần và tiến hành kiểm tra chéo *(cross-validation)*. Hiệu năng của các mô hình trích xuất khía cạnh được đo bởi độ chính xác *(precision)*, độ bao phủ *(recall)* và độ đo F_1 trên mỗi loại danh mục khía cạnh. Lấy ví dụ với danh mục khía cạnh DRINKS#QUALITY (chất lượng đồ uống). Giả sử *A* ký hiệu cho tập các câu có danh mục khía cạnh DRINKS #QUALITY 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 #QUALITY sẽ được tính như sau (tương tự cho các loại danh mục khía cạnh khác):

$$Precision = \frac{|A \cap B|}{|A|} \quad (7)$$

$$Recall = \frac{|A \cap B|}{|B|} \quad (8)$$

và

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

B. *Kết quả thực nghiệm*

Mục đích xây dựng các thực nghiệm:

- Giải quyết nhiệm vụ trích xuất danh mục khía cạnh sử dụng phương pháp học sâu dựa trên kiến trúc BERT.
- Thử nghiệm phương pháp đề xuất trên tập dữ liệu tiếng Việt với các biến thể của hàm mất mát.
- Thử nghiệm phương pháp đề xuất trên tập dữ liệu gồm cả tiếng Việt và tiếng Anh.
- So sánh kết quả của phương pháp đề xuất với kết quả đã thực hiện trong nghiên cứu trước [4].

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

1) ***Giải quyết nhiệm vụ trích xuất danh mục khía cạnh sử dụng phương pháp học sâu dựa trên kiến trúc BERT***

Thử nghiệm đầu tiên chúng tôi thực hiện trên tập dữ liệu tiếng Việt, sử dụng phương pháp trích xuất danh mục khía cạnh dựa trên kiến trúc PhoBERT-*base*, dùng hàm mất mát entropy chéo nhị phân (Binary Cross Entropy - BCE) thường được sử dụng trong phân loại văn bản đa nhãn trong xử lý ngôn ngữ tự nhiên [26]. Kết quả trong Bảng III cho thấy, tính trung bình các độ đo trên toàn bộ 12 nhãn danh mục khía cạnh đạt độ chính xác (precision) là 65%, độ bao phủ (recall) là 62% và độ đo F_1 đạt 62%. Kết quả này sẽ được sử dụng làm baseline cho các thử nghiệm về sau.

Bảng III. Kết quả trích xuất danh mục khía cạnh sử dụng PhoBERT-base và hàm mất mát BCE

TT	Danh mục khía cạnh	Pre.	Rec.	F_1
1	RESTAURANT#GENERAL	0.58	0.10	0.17
2	RESTAURANT#PRICES	0.76	0.19	0.31
3	RESTAURANT#MISCELLANEOUS	0.95	0.04	0.07
4	FOOD#QUALITY	0.93	0.61	0.73
5	FOOD#STYLE_OPTIONS	0.77	0.38	0.51
6	FOOD#PRICES	0.81	0.76	0.78
7	DRINKS#QUALITY	0.77	0.59	0.67
8	DRINKS#STYLE_OPTIONS	0.85	0.71	0.78
9	DRINKS#PRICES	0.89	0.51	0.65
10	SERVICE#GENERAL	0.72	0.33	0.45
11	AMBIENCE#GENERAL	0.82	0.55	0.66
12	LOCATION#GENERAL	0.63	0.51	0.56
	Trung bình	**0.65**	**0.62**	**0.62**

2) ***Thử nghiệm phương pháp đề xuất trên tập dữ liệu tiếng Việt với các biến thể của hàm mất mát***

Bảng IV. Kết quả trích xuất danh mục khía cạnh sử dụng PhoBERT-base và các biến thể của hàm focal loss, tính theo độ đo F_1, trên tập dữ liệu tiếng Việt

TT	Danh mục khía cạnh	BCE	FL	*CB*	*DB*
1	RESTAURANT#GENERAL	0.17	0.64	0.72	0.78
2	RESTAURANT#PRICES	0.31	0.74	0.58	0.73
3	RESTAURANT#MISCELLANEOUS	0.07	0.73	0.61	0.63
4	FOOD#QUALITY	0.73	0.82	0.58	0.31
5	FOOD#STYLE_OPTIONS	0.51	0.84	0.72	0.67
6	FOOD#PRICES	0.78	0.74	0.26	0.61
7	DRINKS#QUALITY	0.67	0.42	0.73	0.79
8	DRINKS#STYLE_OPTIONS	0.78	0.49	0.36	0.83
9	DRINKS#PRICES	0.65	0.79	0.78	0.35
10	SERVICE#GENERAL	0.45	0.31	0.63	0.81
11	AMBIENCE#GENERAL	0.66	0.54	0.48	0.61
12	LOCATION#GENERAL	0.56	0.63	0.82	0.84
	Trung bình	**0.62**	**0.72**	**0.71**	**0.76**

Tiếp theo, chúng tôi thử nghiệm thực hiện trên tập dữ liệu tiếng Việt, sử dụng phương pháp trích xuất danh mục khía cạnh dựa trên kiến trúc PhoBERT*base*, dùng các biến thể khác nhau của hàm mất mát focal loss. Kết quả được trình bày trong Bảng IV, tính theo độ đo F_1. Có thể nhận thấy, kết quả sử dụng 3 biến thể của hàm mất mát là FL (focal loss), hàm focal loss cân bằng theo lớp CB (Class-balanced focal loss), và hàm mất mát cân bằng phân phối DB (Distribution-balanced loss) đều đạt độ đo F_1 trên 71%. Trong đó, phương pháp sử dụng hàm mất mát cân bằng phân phối DB đạt độ đo F_1 cao nhất, là 76%.

3) ***Thử nghiệm phương pháp đề xuất trên tập dữ liệu gồm cả tiếng Việt và tiếng Anh***

Bảng V. Kết quả trích xuất danh mục khía cạnh sử dụng PhoBERT-base và hàm mất mát DB, tính theo độ đo F_1, trên tập dữ liệu gồm cả tiếng Việt và tiếng Anh

TT	Danh mục khía cạnh	Tiếng Việt	Tiếng Việt +Tiếng Anh
1	RESTAURANT#GENERAL	0.78	0.65
2	RESTAURANT#PRICES	0.73	0.57
3	RESTAURANT#MISCELLANEOUS	0.63	0.46
4	FOOD#QUALITY	0.31	0.77
5	FOOD#STYLE_OPTIONS	0.67	0.73
6	FOOD#PRICES	0.61	0.87
7	DRINKS#QUALITY	0.79	0.51
8	DRINKS#STYLE_OPTIONS	0.83	0.59
9	DRINKS#PRICES	0.35	0.72
10	SERVICE#GENERAL	0.81	0.76
11	AMBIENCE#GENERAL	0.61	0.57
12	LOCATION#GENERAL	0.84	0.85
	Trung bình	**0.76**	**0.77**

Với việc bổ sung thêm tập dữ liệu tiếng Anh đã được gán nhãn sẵn [10], chúng tôi tiếp tục thực hiện thử nghiệm trên tập dữ liệu bao gồm cả tiếng Việt và tiếng Anh, sử dụng phương pháp trích xuất danh mục khía cạnh dựa trên kiến trúc PhoBERT-*base*, với hàm mất mát DB (đạt được kết quả tốt nhất trong thử nghiệm trước). Kết quả được trình bày trong Bảng V theo độ đo F_1 cho thấy, phương pháp sử dụng dữ liệu bổ sung đạt kết quả tốt hơn khi chỉ sử dụng dữ liệu tiếng Việt. Cụ thể, tính trung bình, phương pháp sử dụng thêm dữ liệu tiếng Anh có độ đo F_1 là 77%, cao hơn khi chỉ sử dụng dữ liệu tiếng Việt là 1%.

4) ***So sánh kết quả của phương pháp đề xuất với kết quả đã thực hiện trong nghiên cứu trước [4]***

Trong nghiên cứu trước của nhóm [4], chúng tôi sử dụng phương pháp học máy truyền thống (Support Vector Machine, SVM) 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. Tập dữ liệu thử nghiệm bao gồm: tập dữ liệu tiếng Việt (tự gán nhãn) và tập dữ liệu tiếng Việt (tự gán nhãn) được bổ sung thêm từ tập dữ liệu tiếng Anh (đã được gán nhãn sẵn, sau đó được dịch ra tiếng Việt bằng công cụ dịch máy tự động). Việc sử dụng thêm tập dữ liệu tiếng Anh đã làm giàu thêm tài nguyên dữ liệu huấn luyện, giúp cải thiện hiệu suất trích xuất danh mục khía cạnh. Việc trích chọn đặc trưng thủ công cần có kiến thức của chuyên gia về lĩnh vực nghiên cứu.

Trong nghiên cứu này, với việc đề xuất sử dụng phương pháp học sâu dựa trên kiến trúc BERT, cùng với việc sử dụng hàm focal loss, đã giúp loại bỏ quá trình trích chọn đặc trưng thủ công khá tốn kém về thời gian và công sức. Tuy nhiên, kết quả trong Bảng VI cho thấy, hiệu năng trích xuất trung bình được cải thiện đáng kể. Với cả 2 thử nghiệm trên 2 tập dữ liệu tiếng Việt, và Tiếng Việt + Tiếng Anh, phương pháp đề xuất đạt kết quả độ đo F_1 lần lượt là 76% và 77%, đều cao hơn nhiều so với phương pháp trước (tăng 5%).

Bảng VI. So sánh kết quả trung bình tính theo độ đo F_1 của phương pháp đề xuất với kết quả nghiên cứu trước [4]

Nghiên cứu trước [4] SVM + đặc trưng thủ công *(F1-score)*		**Phương pháp đề xuất PhoBERT + DB loss *(F1-score)***	
Tiếng Việt	Tiếng Việt +Tiếng Anh	Tiếng Việt	Tiếng Việt +Tiếng Anh
0.71	0.72	0.76	0.77

VI. 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. 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. Ngoài ra, để khắc phục vấn đề mất cân bằng dữ liệu giữa các nhãn lớp trong bài toán trích xuất thông tin khi tiếp cận theo phương pháp phân loại, chúng tôi đề xuất sử dụng focal loss. Kết quả thực nghiệm cho thấy mô hình trích xuất đề xuất có hiệu năng vượt trội hơn, với kết quả trung bình độ đo F_1 cao nhất đạt 77%.

Trong thời gian tới, chúng tôi dự định nghiên cứu các phương pháp và kỹ thuật mới khác để cải tiến hiệu năng nhiệm vụ này, đồng thời có thể áp dụng kết quả trích xuất khía cạnh cho các nhiệm vụ liên quan tiếp theo. Một số định hướng cụ thể như sau: 1) Nghiên cứu phương pháp học sâu cho trích xuất cảm xúc trong câu văn bản có chứa ý kiến/quan điểm về khía cạnh của dịch vụ/sản phẩm cho tiếng Việt. 2) Nghiên cứu phương pháp trích xuất kết hợp đồng thời cả khía cạnh và cảm xúc trong văn bản.

LỜI CẢM ƠN

Nghiên cứu này được hỗ trợ bởi Học viện Công nghệ Bưu chính Viễn thông, Đề tài hỗ trợ học thuật mã số: 01-2022-HV-CNTT1.

TÀI LIỆU THAM KHẢO

[1] 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.

[2] 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.

[3] 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.

[4] N.T.T. Thuy, N.X. Bach, T.M. Phuong. Cross-Language Aspect Extraction for Opinion Mining. In Proceedings of the 10th International Conference on Knowledge and Systems Engineering (KSE 2018), pp. 67-72, 2018.

[5] 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.

[6] S. Brody and N. Elhadad. An Unsupervised Aspect-sentiment Model for Online Reviews. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 804–812, 2010.

[7] Z. Chen, A. Mukherjee, and B. Liu. Aspect Extraction with Automated Prior Knowledge Learning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 347–358, 2014.

[8] 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.

[9] 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.

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

[11] 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. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), pp. 312–317, 2016.

[12] N.T. Duyen, N.X. Bach, and T.M. Phuong. An empirical study on sentiment analysis for Vietnamese. In Proceedings of the International Conference on Advanced Technologies for Communications (ATC), pp. 309–314, 2014.

[13] Q.V. Ha, B.D.N. Hoang, and M.Q. Nghiem. Lifelong Learning for Cross-Domain Vietnamese Sentiment Classification. In Proceedings of the International Conference on Computational Social Networks (CSoNet), pp. 298–308, 2016.

[14] N.X. Bach, and T.M. Phuong. Leveraging user ratings for resource-poor sentiment classification. In Proceedings of the 19th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), pp. 322–331, 2015.

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

[16] V.N. Phu, V.T.N. Chau, V.T.N. Tran, D.N. Duy, and K.L.D. Duy. A valence-totaling model for Vietnamese sentiment classification. Evolving Systems, pp. 1–47, 2017.

[17] H.S. Le, T.L. Van, and T.V. Pham. Aspect analysis for opinion mining of Vietnamese text. In Proceedings of the International Conference on Advanced Computing and Applications (ACOMP), pp. 118–123, 2015.

[18] T.T. Vu, H.T. Pham, C.T. Luu, and Q.T. Ha. A feature-based opinion mining model on product reviews in Vietnamese. Semantic Methods for Knowledge Management and Communication. Studies in Computational Intelligence, Vol. 381, pp. 23–33, 2011.

[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] J. Cohen. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, Vol. 20, No. 1, pp. 37–46, 1960.

[22] 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.

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

[24] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008. 2017.

[25] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988). 2017.

[26] Y. Bengio, A. Courville and P. Vincent. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), pp.1798-1828. 2013.

[27] Cui, Y., Jia, M., Lin, T.Y., Song, Y. and Belongie, S., 2019. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9268-9277).

[28] Wu, T., Huang, Q., Liu, Z., Wang, Y. and Lin, D., 2020, August. Distribution-balanced loss for multi-label classification in long-tailed datasets. In European Conference on Computer Vision (pp. 162-178). Springer, Cham.

ASPECT CATEGORY EXTRACTION USING BERT WITH BALANCED LOSS FUNCTIONS

Abstract: Aspect category extraction is the first task in the aspect-based opinion mining problem. This is a difficult task because users often use different keywords to describe the same aspect or sometimes only words that are implied to refer to the aspect. Supervised machine learning methods are generally considered to be highly accurate, but are often laborious in annotating the training data, especially for new domains. Furthermore, these methods often require expert knowledge to manually extract useful features to the research domain. This paper presents an enhanced method using BERT-based deep learning model to solve and improve performance for aspect category extraction task. The proposed model automatically learn feature representation efficiently from input text data by using BERT. In addition, to overcome the problem of class imbalance, we suggest using balanced loss functions. Experimental results show that the proposed model has superior performance, with the highest average F_1 measure reaching 77%.

Keywords: aspect category extraction, machine learning, deep learning, BERT, balanced loss function.

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.

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.