Tạp chí Khoa học Đại học Huế: Kỹ thuật và Công nghệ; pISSN 2588-1175 | eISSN 2615-9732

Tập 129, Số 2A, 2020, Tr. 15–29; DOI: 10.26459/hueunijtt.v129i1A.5949

MỞ RỘNG TỪ ĐIỂN VIETSENTIWORDNET CHO MIỀN DỮ LIỆU THUỘC LĨNH VỰC DU LỊCH SỬ DỤNG PHƯƠNG PHÁP DỰA TRÊN TỪ VỰNG

Lê Văn Hòa*

Khoa Du lịch – Đại học Huế

Tóm tắt. Khai phá quan điểm giúp xác định hướng quan điểm (tích cực, tiêu cực) của người dùng về một chủ đề, sản phẩm hay dịch vụ. Có một số cách tiếp cận khác nhau về khai phá quan điểm, trong đó phương pháp khai phá quan điểm dựa trên từ vựng là khá phổ biến. Độ chính xác của phương pháp khai phá quan điểm dựa trên từ vựng phụ thuộc rất nhiều vào từ điển được sử dụng, trong đó chứa các từ quan điểm về các lĩnh vực cụ thể. Một bộ dữ liệu có thể thực hiện phân lớp tốt trong lĩnh vực này, nhưng lại kém hiệu quả đối với một số lĩnh vực khác. VietSentiWordNet là từ điển quan điểm tiếng Việt được sử dụng khá phổ biến hiện nay, nhưng thiếu nhiều từ quan điểm cho miền dữ liệu thuộc lĩnh vực du lịch. Bài báo này do đó tập trung vào việc mở rộng từ điển VietSentiWordNet với việc làm giàu các từ quan điểm thuộc lĩnh vực du lịch, trong đó một mô hình mở rộng từ điển VietSentiWordNet sử dụng phương pháp dựa trên từ vựng được đề xuất và một tiến trình tiền xử lý dữ liệu với nhiều chức năng được tích hợp cũng được bổ sung nhằm nâng cao hiệu quả phân lớp quan điểm. Kết quả thực nghiệm cho thấy rằng việc từ điển VietSentiWordNet được mở rộng đã phân lớp quan điểm chính xác hơn đối với các câu quan điểm trong lĩnh vực du lịch.

Từ khóa: Từ điển VietSentiWordNet; miền dữ liệu du lịch; khai phá quan điểm; phương pháp dựa trên từ vựng; hướng quan điểm.

1 Giới thiệu

Khai phá quan điểm là một phân nhánh khác của xử lý ngôn ngữ tự nhiên, nghiên cứu về thái độ, cảm xúc hoặc tình cảm của con người về một chủ đề, sản phẩm, hàng hóa hoặc dịch vụ cụ thể [1]. Lĩnh vực này liên quan đến xử lý ngôn ngữ, ngôn ngữ học tính toán kết hợp với khai phá văn bản, bao gồm các kỹ thuật liên quan đến khai phá dữ liệu về quan điểm và cố gắng thu thập thông tin có giá trị từ dữ liệu này. Theo W. Medhat và cộng sự [2], khai phá quan điểm có thể được tiếp cận theo 3 hướng chính: Học máy (machine-learning), dựa trên từ vựng (lexicon-based), và lai (kết hợp hai phương pháp trên). Phương pháp dựa trên từ vựng thực hiện việc tìm kiếm các từ vựng về quan điểm để phân tích văn bản. Từ vựng về quan điểm tích cực thể hiện một số trạng thái mong muốn (như: yêu, thích, …), trong khi từ vựng về quan điểm tiêu cực thể hiện một số trạng thái không mong muốn (như: ghét, không thích, …). Ngoài ra, còn có các cụm từ vựng về quan điểm và thành ngữ mà sau đây được gọi chung là từ vựng quan điểm.

* *Liên hệ: lvhoa@hueuni.edu.vn*

Nhận bài: 03–8–2020; Hoàn thành phản biện: 07–9–2020; Ngày nhận đăng: 08–9–2020

Khai phá quan điểm dựa trên từ vựng thực tế cũng đã thu hút được nhiều nhà khoa học quan tâm. Cụ thể, nghiên cứu của C. Bucur [3] đã xây dựng một mô hình khai phá quan điểm, trích xuất các đánh giá về khách sạn từ các diễn đàn trên Internet và phân loại chúng dựa vào từ điển SentiWordNet [4]; V. Soni và cộng sự [5] sử dụng phương pháp dựa trên từ vựng kết hợp từ điển SentiWordNet để tìm ra các khía cạnh tích cực và tiêu cực của sản phẩm điện thoại di động trên website Amazon.com; M. Kundi và cộng sự [6] đã đề xuất một mô hình sử dụng phương pháp dựa trên từ vựng để phân lớp quan điểm với dữ liệu là các tweet trên mạng xã hội Twitter và G. Qiu và cộng sự [7] đã sử dụng phương pháp dựa trên từ điển để xác định các câu quan điểm trong quảng cáo theo ngữ cảnh.

Đối với tiếng Việt, nghiên cứu của Kiều Thanh Bình và cộng sự [8] sử dụng từ điển liên quan đến các đặc trưng về cấu hình và kiểu dáng máy tính. Vũ Tiến Thành và cộng sự [9] đã xây dựng mô hình khai phá quan điểm khách hàng về các sản phẩm điện thoại di động dựa vào luật cú pháp tiếng Việt và từ điển VietSentiWordNet [10]. Rõ ràng, tùy thuộc vào từng lĩnh vực ứng dụng mà các nghiên cứu này làm giàu thêm các từ quan điểm cho lĩnh vực đó và kết quả là các mô hình khai phá quan điểm dựa trên các từ điển mở rộng này đã nâng cao được hiệu quả phân lớp quan điểm. Riêng với lĩnh vực du lịch, chưa có nghiên cứu nào về khai phá quan điểm đối với miền dữ liệu tiếng Việt.

Theo P. Haseena Rahmath [11], thách thức lớn nhất đối với khai phá quan điểm là đặc tính phụ thuộc lĩnh vực của các từ quan điểm. Một bộ dữ liệu tại cùng một thời điểm có thể thực hiện phân lớp tốt trong lĩnh vực này trong khi thực hiện phân lớp kém hiệu quả đối với các lĩnh vực khác. Cùng chung với quan điểm này, Hong Nam Nguyen và cộng sự [12] cho rằng những từ điển quan điểm đang tồn tại một số giới hạn nhất định khi áp dụng để phân tích các bình luận và đánh giá tiếng Việt trong khai phá quan điểm người sử dụng. Đa số các từ điển được sử dụng trong các mô hình khai phá dữ liệu này thiếu khá nhiều từ quan điểm, đặc biệt trong các lĩnh vực cụ thể, dẫn đến hiệu quả phân lớp không cao. Từ những lý do đó, chúng tôi đề xuất mở rộng từ điển VietSentiWordNet của Vũ Xuân Sơn và cộng sự [10] với việc làm giàu thêm các từ quan điểm liên quan đến lĩnh vực du lịch. Để thực hiện điều đó, chúng tôi đề xuất một mô hình mở rộng từ điển VietSentiWordNet cho miền dữ liệu thuộc lĩnh vực du lịch sử dụng phương pháp dựa trên từ vựng. Ý tưởng xây dựng mô hình này xuất phát từ các nghiên cứu trong [6, 8, 13, 14] với dữ liệu vào là các bình luận liên quan đến lĩnh vực du lịch. Thêm vào đó, chúng tôi đề xuất một tiến trình tiền xử lý dữ liệu với một số chức năng tích hợp nhằm nâng cao hiệu quả phân lớp, như thêm dấu, chuẩn hóa láy âm tiết (đối với những từ thể hiện cảm xúc đặc biệt), chuẩn hóa chữ viết tắt, xử lý biểu tượng cảm xúc. Các đề xuất này nhằm hướng đến xây dựng một từ điển quan điểm tiếng Việt mà có thể áp dụng cho việc phân lớp quan điểm trong lĩnh vực du lịch.

Các phần tiếp theo của bài báo gồm: các phân tích về các nghiên cứu liên quan được mô tả trong Phần 2; Phần 3 đề xuất phương pháp mở rộng từ điển VietSentiWordNet cho miền dữ liệu du lịch, trong đó hai sơ đồ bổ sung từ quan điểm và tiền xử lý dữ liệu được mô tả chi tiết. Phần 4 là thực nghiệm và phân tích kết quả. Kết luận của bài báo được trình bày trong Phần 5.

2 Nghiên cứu liên quan

Khai phá quan điểm là nhằm phát hiện quan điểm về một đối tượng là tích cực hay tiêu cực. Các đặc trưng về đối tượng là được mô tả, đánh giá ở các mức độ khác nhau. Theo B. Liu [15], các thành phần cơ bản của một quan điểm bao gồm:

- Người nêu quan điểm (Opinion holder): là người hoặc tổ chức đưa ra quan điểm về một đối tượng.
- Đối tượng (Object): là một thực thể được phản ảnh bởi người nêu quan điểm đưa ra quan điểm.
- Quan điểm (Opinion): là một ý kiến, tình cảm hoặc sự đánh giá của người nêu quan điểm về một đối tượng.

Tùy theo từng trường hợp và mục đích cụ thể, việc khai phá quan điểm có thể ở các mức khác nhau: mức tài liệu, mức câu và mức đặc trưng. Dựa vào nhiệm vụ liên quan đến các mức và giả định được thực hiện ở các mức khác nhau, N. Mishra và cộng sự [16] đã đưa ra đánh giá về khai phá quan điểm ở các mức khác nhau như được mô tả như trong Bảng 1.

Bảng 1. Đánh giá về khai phá quan điểm ở các mức khác nhau

Mức khai phá quan điểm	Giả định được thực hiện	Nhiệm vụ liên quan
Mức tài liệu	1. Mỗi tài liệu tập trung vào một đối tượng và chứa quan điểm được đưa ra bởi duy nhất một người nêu quan điểm. 2. Không áp dụng cho bài đăng trên blog và diễn đàn vì có thể có nhiều quan điểm về nhiều đối tượng trong các nguồn đó.	Nhiệm vụ: Phân loại quan điểm đánh giá. Các lớp: Tích cực, tiêu cực và trung lập.
Mức câu	1. Một câu chỉ chứa duy nhất một quan điểm được đăng bởi duy nhất một người nêu quan điểm; điều này không thể đúng trong nhiều trường hợp, ví dụ có thể có nhiều quan điểm trong câu ghép và câu phức. 2. Ranh giới câu được xác định trong tài liệu đã cho.	Nhiệm vụ 1: Xác định câu đã cho là chủ quan hoặc có quan điểm. Các lớp: Khách quan và chủ quan (có quan điểm). Nhiệm vụ 2: Phân loại quan điểm của câu đã cho. Các lớp: Tích cực, tiêu cực và trung lập.

Mức khai phá quan điểm	Giả định được thực hiện	Nhiệm vụ liên quan
Mức đặc trưng	1. Nguồn dữ liệu tập trung vào các đặc trưng của một đối tượng được đăng bởi duy nhất người nêu quan điểm. 2. Không áp dụng cho bài đăng trên blog và diễn đàn vì có thể có nhiều quan điểm về nhiều đối tượng trong các nguồn đó.	Nhiệm vụ 1: Xác định và trích xuất các đặc trưng đối tượng đã được nhận xét bởi người nêu quan điểm. Nhiệm vụ 2: Xác định hướng các quan điểm về các đặc trưng là tích cực, tiêu cực hay trung lập. Nhiệm vụ 3: Gom nhóm đặc trưng đồng nghĩa. Tạo một bản tóm tắt quan điểm dựa trên đặc trưng của nhiều đánh giá.

Đã có một số nghiên cứu liên quan đến khai phá quan điểm sử dụng phương pháp dựa trên từ vựng. Cụ thể, nghiên cứu của C. Bucur [3] đã đề xuất một mô hình để trích xuất và phân loại các đánh giá khách sạn được đăng bởi người dùng trên các website du lịch. Hệ thống trích xuất các đánh giá của khách sạn từ internet và sử dụng kỹ thuật khai phá quan điểm để phân loại chúng dựa vào từ điển SentiWordNet [4]. Tuy nhiên, mô hình khai phá quan điểm của tác giả xử lý ở nhiều mức (mức từ, mức câu và mức tài liệu) nên sẽ gặp khó khăn khi tổng hợp quan điểm đối với các bình luận chứa nhiều quan điểm liên quan đến nhiều đối tượng. Tương tự, nghiên cứu của V. Soni và cộng sự [5] cũng đã sử dụng phương pháp dựa trên từ vựng kết hợp với từ điển SentiWordNet. Nhóm tác giả tập trung vào việc phân tích quan điểm ở cấp độ khía cạnh để tìm ra các khía cạnh tích cực và tiêu cực của sản phẩm điện thoại trên website Amazon.com. Mục tiêu chính của phân tích mức khía cạnh là xác định các đặc trưng sẽ được phân tích, trích xuất các đặc trưng này và tính toán độ phân cực của nó. Trong nghiên cứu này, nhóm tác giả đã thiết kế một bộ từ điển dữ liệu mới cho lĩnh vực điện thoại di động. Tuy nhiên, trong giai đoạn tiền xử lý dữ liệu của nhóm tác giả chưa tích hợp các chức năng xử lý chữ viết tắt và biểu tượng cảm xúc để tăng ngữ nghĩa cho văn bản. Trong khi đó, nghiên cứu của M. Kundi và cộng sự [6] đã đề xuất một mô hình sử dụng phương pháp dựa trên từ vựng để phân lớp quan điểm với dữ liệu là các tweet trên mạng xã hội Twitter. Mô hình này dựa trên sự tổng hợp của nguồn dữ liệu là các bộ từ vựng và từ điển khác nhau. Nghiên cứu này quan tâm đến việc xử lý tiếng lóng và biểu tượng cảm xúc giúp cho việc phân lớp đạt hiệu quả cao. Ngoài ra, nghiên cứu của G. Qiu và cộng sự [7] đã sử dụng phương pháp dựa trên từ điển để xác định các câu quan điểm trong quảng cáo theo ngữ cảnh. Nhóm tác giả đã đề xuất một chiến lược quảng cáo để cải thiện mức độ phù hợp của quảng cáo và trải nghiệm người dùng. Nhóm tác giả cũng đã sử dụng phân tích cú pháp và từ điển quan điểm sau đó đề xuất một cách tiếp cận dựa trên các luật để giải quyết vấn đề trích xuất chủ đề và nhận dạng thái độ của người tiêu dùng trong trích xuất từ khóa quảng cáo. Kết quả của nhóm tác giả đã chứng minh tính hiệu quả của phương pháp đề xuất về trích xuất từ khóa quảng cáo và lựa chọn quảng cáo. Tuy nhiên, nhóm tác giả chỉ quan tâm đến việc trích xuất

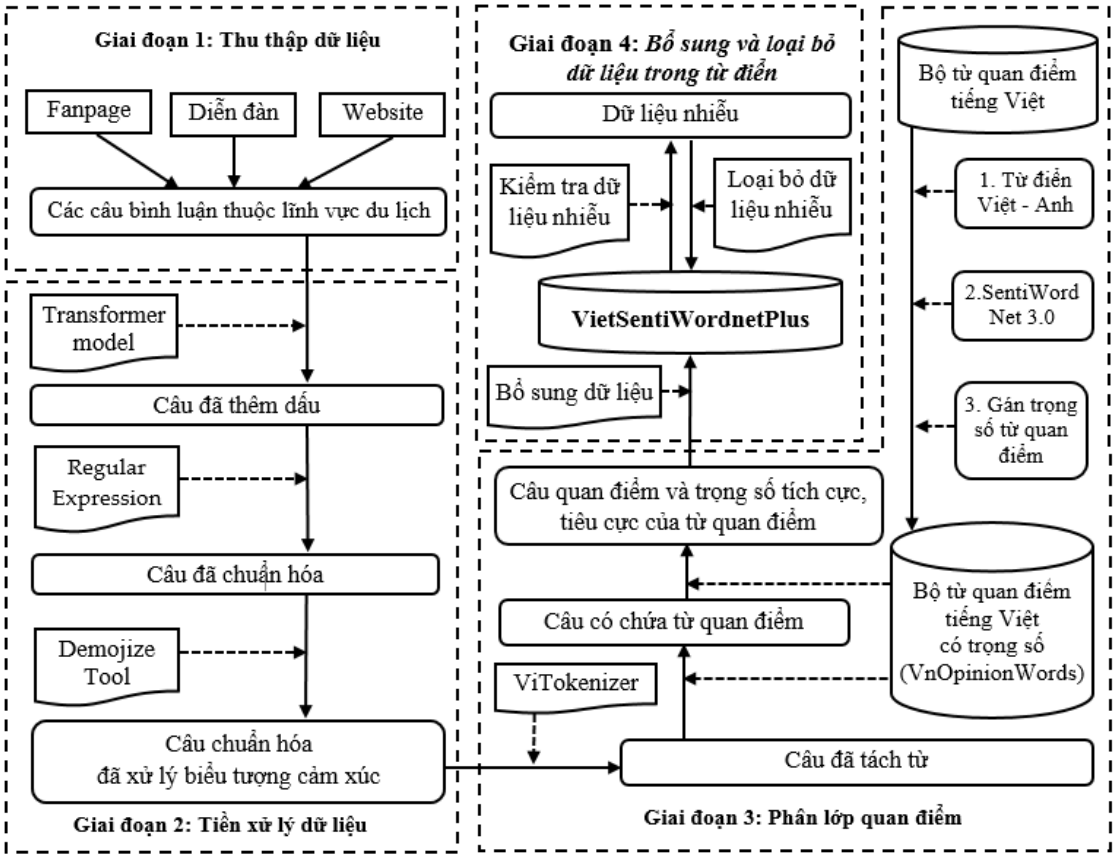
từ khóa quảng cáo mà chưa tích hợp chức năng tiền xử lý dữ liệu để tăng ngữ nghĩa cho dữ liệu phân tích.

Ở trong nước, nghiên cứu của Kiều Thanh Bình và Phạm Bảo Sơn [8] đã đề xuất hệ thống khai phá quan điểm cho sản phẩm máy tính từ các bình luận tiếng Việt sử dụng phương pháp dựa trên luật để xây dựng các đánh giá tự động quan điểm của người dùng ở mức câu, trong đó nhóm tác giả đã sử dụng các từ điển dữ liệu khác nhau để phân lớp các đặc trưng và quan điểm như từ điển các từ liên quan đến các đặc trưng cấu hình của máy tính (cấu hình, vi xử lý, hệ thống...); từ điển các từ liên quan đến các đặc trưng kiểu dáng (thiết kế, thân máy, màu sắc, kích cỡ...); từ điển chứa các từ quan điểm (tích cực, tiêu cực). Tuy nhiên, quá trình tiền xử lý dữ liệu của nhóm tác giả chỉ quan tâm đến việc tách từ, gán nhãn từ loại cho câu nhưng chưa quan tâm đến việc chuẩn hóa dữ liệu đối với các trường hợp dữ liệu tiếng Việt không dấu hoặc có chứa biểu tượng cảm xúc nên hiệu quả phân lớp không cao. Trong khi đó, nghiên cứu của Vũ Tiến Thành và cộng sự [9] đã trình bày một phương pháp xây dựng hệ thống khai phá quan điểm của khách hàng về đặc trưng của sản phẩm dựa vào luật cú pháp tiếng Việt và từ điển VietSentiWordNet. Kết quả, nhóm tác giả xây dựng mô hình khai phá và tổng hợp quan điểm dựa trên đặc trưng từ các đánh giá của khách hàng đối với sản phẩm điện thoại di động. Điểm hạn chế của mô hình đề xuất là nhóm tác giả chưa quan tâm đến việc xử lý biểu tượng cảm xúc và chữ viết tắt trong pha thứ nhất (pha tiền xử lý dữ liệu). Ngoài ra, Hong Nam Nguyen và cộng sự [12] đã đề xuất phương pháp khai phá quan điểm văn bản tiếng Việt sử dụng từ điển quan điểm cho miền cụ thể để tăng độ chính xác. Từ điển quan điểm được nhóm tác giả xây dựng quan tâm đến miền dữ liệu là các sản phẩm điện tử như điện thoại di động và máy tính. Để nâng cao hiệu quả phân lớp, nhóm tác giả đã xử lý được trường hợp câu bình luận chứa chữ viết tắt, tiếng lóng. Mô hình của nhóm tác giả đề xuất phụ thuộc vào độ chính xác của phần mềm dịch từ tiếng Anh sang tiếng Việt, từ tiếng Việt sang tiếng Anh và từ điển Việt – Việt, đây cũng là thách thức rất lớn ảnh hưởng đến độ chính xác của hệ thống. Từ các nghiên cứu liên quan ở trên, chúng tôi thấy rằng các nghiên cứu này chưa quan tâm đến từ điển dữ liệu cho miền dữ liệu thuộc lĩnh vực du lịch. Ngoài ra, một số nghiên cứu chưa quan tâm đến các chức năng tiền xử lý dữ liệu giúp cho việc phân lớp có độ chính xác và đạt hiệu quả cao. Với đặc thù nguồn dữ liệu là các câu quan điểm được thu thập từ các bình luận, ý kiến đánh giá của khách hàng, và quá trình xây dựng bộ từ điển dữ liệu chưa cần quan tâm đến đặc trưng của đối tượng nghiên cứu mà chỉ quan tâm đến hướng quan điểm (tích cực, tiêu cực) của câu nên chúng tôi chọn phương pháp khai phá quan điểm dựa trên từ vựng ở mức câu là phù hợp với bài toán xây dựng bộ từ điển.

3 Mô hình mở rộng từ điển VietSentiWordNet cho miền dữ liệu du lịch

Chúng tôi mở rộng từ điển VietSentiWordNet với nhiều bổ sung liên quan đến các từ quan điểm thuộc lĩnh vực du lịch dựa vào các nghiên cứu trong [6, 8, 13, 14] là các mô hình phân lớp

quan điểm sử dụng phương pháp dựa trên từ vựng. Hình 1 mô tả mô hình mở rộng từ điển VietSentiWordNet cho miền dữ liệu thuộc lĩnh vực du lịch sử dụng phương pháp dựa trên từ vựng. Mô hình bao gồm 4 giai đoạn thực hiện như sau: (1) Thu thập dữ liệu (2) Tiền xử lý dữ liệu (3) Phân lớp quan điểm (4) Bổ sung và loại bỏ dữ liệu trong từ điển.



Hình 1. Mô hình mở rộng từ điển VietSentiWordNet cho miền dữ liệu thuộc lĩnh vực du lịch

3.1 Giai đoạn 1: Thu thập dữ liệu

Để có dữ liệu đưa vào phân tích, chúng tôi tiến hành thu thập các bình luận, ý kiến đánh giá từ nhiều nguồn khác nhau như các Fanpage, diễn đàn, website... liên quan đến khách sạn, nhà hàng, điểm du lịch, công ty lữ hành. Với mục tiêu thu thập được dữ liệu từ nhiều nguồn khác nhau, quá trình thu thập dữ liệu được thực hiện tự động đối với nguồn dữ liệu là các Fanpage điểm du lịch và bán tự động đối với các nguồn dữ liệu là các website, diễn đàn. Bảng 2 liệt kê hơn 31.630 câu bình luận thu thập được từ nhiều nguồn dữ liệu khác nhau. Từ dữ liệu sau khi thu thập, chúng tôi tách ra làm 2 bộ dữ liệu (bộ dữ liệu đầu vào cho quá trình xây dựng từ điển dữ liệu và bộ dữ liệu kiểm thử).

Bảng 2. Số câu bình luận thu thập được từ nhiều nguồn dữ liệu khác nhau

STT	Địa chỉ	Lĩnh vực	Số câu bình luận
1.	https://facebook.com/diemdulichhue	Điểm du lịch	476
2.	https://facebook.com/assessdestination	Điểm du lịch	768
3.	https://facebook.com/khamphadiemdulichhue	Điểm du lịch	1.058
4.	https://facebook.com/Đại-Nội-Huế-Imperial-City-Hue-356344234958914	Điểm du lịch	1.131
5.	https://traveloka.com	Khách sạn	698
6.	https://tripadvisor.com.vn	Nhà hàng	640
7.	https://liberzy.com	Tour du lịch	122
8.	https://tripadvisor.com.vn	Tour du lịch	819
9.	https://www.shopee.vn	Sản phẩm	567
10.	https://www.shopee.vn	Du lịch	241
11.	https://www.foody.vn/hue (Điểm du lịch di tích, nghỉ dưỡng, sinh thái, tham quan)	Điểm du lịch tại Huế	1.575
12.	https://www.foody.vn/da-nang (Điểm du lịch sinh thái, tham quan, nghỉ dưỡng, khu vui chơi)	Điểm du lịch tại Đà Nẵng	2.501
13.	https://www.tripadvisor.com.vn/Attractions-g293926-Activities-Hue_Thua_Thien_Hue_Province.html (Điểm di tích lịch sử, tôn giáo, hoạt động ngoài trời, tham quan ẩm thực)	Hoạt động giải trí tại Huế	8.331
14.	https://www.tripadvisor.com.vn/Attractions-g298085-Activities-Da_Nang.html (Điểm du lịch thiên nhiên, danh lam thắng cảnh, bảo tàng, siêu thị)	Hoạt động giải trí tại Đà Nẵng	4.732
15.	https://www.tripadvisor.com.vn/Attractions-g293925-Activities-c57-Ho_Chi_Minh_City.html (Điểm du lịch thiên nhiên, danh lam thắng cảnh, hoạt động ngoài trời)	Hoạt động giải trí tại TPHCM	7.975

3.2 Giai đoạn 2: Tiền xử lý dữ liệu

Dữ liệu đầu vào của giai đoạn này là các câu bình luận đã thu thập được. Để nâng cao hiệu quả phân lớp và chất lượng dữ liệu của từ điển, chúng tôi dựa vào nghiên cứu [6, 17] để tích hợp các chức năng tiền xử lý dữ liệu bao gồm thêm dấu, chuẩn hóa láy âm tiết, chuẩn hóa chữ viết

tắt, xử lý biểu tượng cảm xúc. Trong chức năng tiền xử lý dữ liệu đầu tiên, chúng tôi tiến hành thêm dấu cho câu bình luận đối với các câu tiếng Việt không dấu. Vấn đề thêm dấu được đưa về bài toán dịch máy trong đó ngôn ngữ nguồn là tiếng Việt không dấu và ngôn ngữ đích là tiếng Việt có dấu. Bài toán dịch máy cụ thể là Sequence-to-Sequence Learning với kiến trúc Encoder-Decoder đạt hiệu quả cao khi sử dụng mô hình Transformer [18]. Trong chức năng tiền xử lý dữ liệu tiếp theo, chúng tôi tiến hành chuẩn hóa dữ liệu tiếng Việt sử dụng các quy tắc trong biểu thức chính quy (Regular Expression). Trường hợp thứ nhất: chuẩn hóa láy âm tiết (đối với những từ thể hiện cảm xúc đặc biệt), ví dụ: câu bình luận *"Chất lượng dịch vụ tuyệt vòiiiiiiii"* sẽ được chuẩn hóa thành *"Chất lượng dịch vụ tuyệt vời"* hoặc "*Thức ăn ngonnn quá điiiiiiii !!!!!!!!*" sẽ được chuẩn hóa thành "*Thức ăn ngon quá đi !*". Trường hợp thứ hai: chuẩn hóa chữ viết tắt, hệ thống thực hiện việc thay thế các từ như: *"ko"*, *"khong"* thành từ *"không"* hoặc *"đc"*, *"dc"* thành từ *"được"* hay *"ok"*, *"nice"*, *"good"* thành từ *"tốt"* để nâng cao hiệu quả xác định hướng quan điểm cho các câu bình luận. Ngoài ra, chúng tôi còn dựa vào công cụ Demojize [19] để xử lý biểu tượng cảm xúc bằng cách chuyển các biểu tượng cảm xúc này thành văn bản. Bảng 3 mô tả danh sách các biểu tượng cảm xúc được chuyển sang dạng văn bản theo quy định của công cụ Demojize. Kết thúc giai đoạn này, chúng tôi thu thập được các câu bình luận đã chuẩn hóa và xử lý biểu tượng cảm xúc.

Bảng 3. Danh sách các biểu tượng cảm xúc được chuyển sang dạng văn bản

STT	Biểu tượng	Dạng văn bản
1.	😠	angry_face
2.	😧	anguished_face
3.	💔	broken_heart
4.	🥶	cold_face
5.	😘	face_blowing_a_kiss
6.	😬	grimacing_face
7.	😀	grinning_face

STT	Biểu tượng	Dạng văn bản
8.	😭	loudly_crying_face
9.	😔	pensive_face
10.	❤️	red_heart
…		
103.	☺️	smiling_face
104.	👎	thumbs_down
105.	👍	thumbs_up

3.3 Giai đoạn 3: Phân lớp quan điểm

Dữ liệu đầu vào của giai đoạn này là các câu bình luận đã qua xử lý. Bước đầu tiên, chúng tôi dựa vào công cụ ViTokenizer [20] để thực hiện tách từ trong câu. Công cụ ViTokenizer sử

dụng thuật toán Conditional Random Field với độ chính xác hơn 98,50% cho tách từ tiếng Việt. Bước tiếp theo của giai đoạn này, chúng tôi dựa vào nghiên cứu [13, 14] để xây dựng quy trình phân lớp quan điểm gồm 2 công việc chính: xây dựng bộ từ quan điểm tiếng Việt có trọng số (VnOpinionWords) và dựa vào bộ từ quan điểm này để xác định câu quan điểm với trọng số tích cực, tiêu cực của từ quan điểm. Công việc đầu tiên, chúng tôi xây dựng bộ từ quan điểm tiếng Việt có trọng số VnOpinionWords chứa các từ quan điểm và trọng số (tích cực, tiêu cực) của các từ quan điểm. Để xây dựng bộ từ quan điểm này, chúng tôi sử dụng từ điển Việt – Anh để dịch bộ từ quan điểm tiếng Việt sang tiếng Anh, sau đó gán trọng số (tích cực, tiêu cực) cho các từ quan điểm tiếng Việt dựa vào trọng số của các từ tiếng Anh tương ứng trong từ điển SentiWordNet 3.0 [4]. SentiWordNet 3.0 là nguồn từ vựng được tạo ra để hỗ trợ các ứng dụng khai phá quan điểm với ngôn ngữ tiếng Anh. Công việc tiếp theo là xác định câu quan điểm với trọng số tích cực, tiêu cực của từ quan điểm, chúng tôi sử dụng phương pháp dựa trên từ vựng kết hợp với bộ từ quan điểm tiếng Việt có trọng số VnOpinionWords để xác định câu có chứa từ quan điểm, sau đó tính trọng số tích cực, tiêu cực cho từ quan điểm trong câu quan điểm.

3.4 Giai đoạn 4: Bổ sung và loại bỏ dữ liệu trong từ điển

Dữ liệu của từ điển VietSentiWordNetPlus được mở rộng từ từ điển VietSentiWordNet của Vũ Xuân Sơn và cộng sự [10] với khoảng 900 tập từ quan điểm. Hệ thống tự động bổ sung từ quan điểm vào bộ từ điển dữ liệu này dựa vào kết quả phân lớp quan điểm câu bình luận ở giai đoạn 3 (phân lớp quan điểm). Để đảm bảo dữ liệu trong từ điển không trùng lặp, hệ thống kiểm tra sự tồn tại của từ quan điểm trong bộ từ điển, sau đó bổ sung dữ liệu vào từ điển theo đúng khuôn dạng được mô tả như trong Bảng 4. Quá trình chạy thực nghiệm đã bổ sung thêm hơn 1,710 từ quan điểm thuộc lĩnh vực du lịch vào từ điển VietSentiWordNetPlus. Như vậy, số lượng từ quan điểm thuộc lĩnh vực du lịch được bổ sung vào từ điển VietSentiWordNetPlus lớn hơn gần gấp hai lần (từ 900 lên 2,615) số từ quan điểm đã có trong từ điển VietSentiWordNet.

Bảng 4. Khuôn dạng từ quan điểm trong từ điển dữ liệu

STT	PosScore	NegScore	SynsetTerms	Gloss
1.	0,5	0	trong_lành	Không khí trong lành
2.	0,625	0	tuyệt	Cảnh vật đẹp tuyệt
3.	0	0,125	chật_hẹp	Không gian chật hẹp lắm
4.	0,75	0	hùng_vĩ	Phong cảnh hùng vĩ
5.	0	0,875	nghèo_nàn	Thức ăn sáng nghèo nàn
…				

2612.	1	0	:relieved_face	Biển đẹp, đồ ăn lại ngon nữa chứ 😌😌
2613.	0	0,625	nguy_hiểm	Trời mưa đi nguy hiểm
2614.	0	0,875	lộn_xộn	Biển dạo này đông đúc và lộn xộn lắm
2615.	0	0,625	gồ_ghề	Đường kiệt vào khách sạn khá gồ ghề

Ngoài ra, để nâng cao hiệu quả phân lớp của dữ liệu trong từ điển dữ liệu, chúng tôi thực hiện giai đoạn loại bỏ dữ liệu nhiễu. Mục đích của giai đoạn này nhằm loại bỏ những từ quan điểm trong từ điển phân lớp không chính xác thuộc lĩnh vực du lịch. Trong quá trình chạy thử nghiệm từ điển VietSentiWordNet ban đầu, chúng tôi đã phát hiện 12 từ quan điểm phân lớp không chính xác (câu tích cực mà hệ thống cho là câu tiêu cực). Bảng 5 mô tả danh sách các từ quan điểm phân lớp không chính xác. Bộ từ điển VietSentiWordNetPlus đã khắc phục được hạn chế này giúp cho kết quả phân lớp chính xác hơn.

Bảng 5. Danh sách các từ quan điểm phân lớp không chính xác

STT	Từ quan điểm	Ví dụ câu bình luận phân lớp không chính xác
1.	cho	Thuận tiện cho việc di chuyển; Địa điểm lí tưởng cho du lịch
2.	lành	Không khí rất trong lành
3.	sống	Chỗ này sống ảo thì tuyệt vời
4.	mát mẻ	Không khí trong lành mát mẻ
	....	
11.	phong phú	Mặt hàng phong phú; Kiến trúc phong phú
12.	xanh	Vườn cây xanh mát; Biển đẹp và xanh

4 Thực nghiệm và phân tích kết quả

Trong thực nghiệm, có rất nhiều độ đo được sử dụng để đánh giá hiệu suất của bộ phân loại. Trong đó, bốn độ đo được sử dụng rộng rãi bao gồm: Accuracy, Precision, Recall, và F1-score [21]. Ngoài ra, ma trận Confusion là một công cụ rất hữu ích giúp phân tích mức độ hiệu quả mà bộ

phân loại có thể phân loại các mẫu dữ liệu của các lớp khác nhau. Ví dụ về các tham số của ma trận Confusion đối với hai lớp tích cực, tiêu cực được minh họa như trong Bảng 6.

Bảng 6. Ma trận Confusion đối với hai lớp tích cực, tiêu cực

		Mẫu dữ liệu thực tế	
		Tích cực (Positive)	Tiêu cực (Negative)
Bộ phân loại	Tích cực (Positive)	True Positive (TP)	False Positive (FP)
	Tiêu cực (Negative)	False Negative (FN)	True Negative (TN)

Ý nghĩa các tham số trong ma trận Confusion đối với hai lớp tích cực, tiêu cực:

- True Positive (TP): số mẫu của lớp Positive được bộ phân loại dự đoán chính xác là Positive.
- True Negative (TN): số mẫu của lớp Negative được bộ phân loại dự đoán chính xác là Negative.
- False Positive (FP): số mẫu của lớp Negative bị bộ phân loại dự đoán nhầm thành Positive.
- False Negative (FN): số mẫu của lớp Positive bị bộ phân loại dự đoán nhầm thành Negative.

Một số độ đo đánh giá hiệu suất của bộ phân loại:

Độ chính xác tổng quát (Accuracy) xác định hiệu suất của bộ phân loại là tỷ lệ phần trăm mẫu được dự đoán chính xác. Accuracy được tính bằng tỷ số giữa số mẫu được dự đoán chính xác (không phân biệt Positive hay Negative) trên tổng số mẫu. Công thức tính độ chính xác tổng quát (Accuracy):

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

Độ chính xác (Precision) cho biết độ chính xác của bộ phân loại là tỷ lệ phần trăm của tất cả các mẫu được dự đoán tính cực thực sự là tính cực. Công thức tính độ chính xác (Precision):

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

Độ đầy đủ (Recall) thường liên quan đến thước đo mức độ đầy đủ của bộ phân loại là tỷ lệ phần trăm mẫu tích cực thực sự được dự đoán chính xác là tích cực. Công thức tính độ đầy đủ (Recall):

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

Độ đầy đủ điều hòa (F-score) là một thước đo phân tích thống kê có tính đến cả độ chính xác và mức độ đầy đủ, F-score có giá trị từ 0 đến 1. Giá trị F-score càng gần với 1 thì độ chính xác của bộ phân loại càng cao. Công thức tính độ đầy đủ điều hòa (F-score):

$$\text{F-score} = 2\frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Một ví dụ minh họa về kết quả đánh giá thực nghiệm của hệ thống phân lớp sử dụng từ điển VietSentiWordNetPlus đối với điểm du lịch Đại Nội Huế được mô tả như trong Bảng 7.

Bảng 7. Kết quả đánh giá thực nghiệm đối với điểm du lịch Đại Nội Huế

		Mẫu dữ liệu thực tế		Kết quả đánh giá			
		Positive	Negative	Accuracy	Precision	Recall	F-score
Hệ thống phân lớp	Positive	TP là 92	FP là 08	90,58 %	92,00 %	94,85 %	93,40 %
	Negative	FN là 05	TN là 33		86,84 %	80,49 %	83,54 %
Trung bình					89,42 %	87,67 %	88,47 %

Từ kết quả đánh giá thực nghiệm trong Bảng 7, chúng ta thấy hiệu suất phân lớp sử dụng từ điển VietSentiWordNetPlus đối với điểm du lịch Đại Nội Huế là khá cao. Trong đó, hệ thống dự đoán một lớp là Positive có Precision(Positive) là 92,00%, Recall(Positive) là 94,85%, F-score(Positive) là 93,40%; dự đoán một lớp là Negative có Precision(Negative) là 86,84%, Recall(Negative) là 80,49%, F-score(Negative) là 83,54%; độ chính xác tổng quát (Accuracy) là 90,58%. Như vậy, hiệu suất phân lớp trung bình sử dụng từ điển VietSentiWordNetPlus đối với điểm du lịch Đại Nội Huế có độ chính xác tổng quát (Accuracy) là 90,58%, độ chính xác (Precision) là 89,42%, độ đầy đủ (Recall) là 87,67% và độ đầy đủ điều hòa (F-score) là 88,47%.

Chúng tôi thực hiện cài đặt với các thiết lập tương tự như trong hệ thống phân lớp quan điểm của Vũ Xuân Sơn và cộng sự để so sánh kết quả thực nghiệm phân lớp quan điểm giữa bộ từ điển VietSentiWordNet và bộ từ điển VietSentiWordNetPlus (của chúng tôi mở rộng). Bảng 8 so sánh hiệu suất xác định hướng quan điểm (theo phương pháp Accuracy và Precision - Recall) của 10 điểm du lịch giữa từ điển VietSentiWordNetPlus với từ điển VietSentiWordNet. Kết quả đánh giá hiệu suất trung bình xác định hướng quan điểm của bộ từ điển VietSentiWordNetPlus về độ chính xác tổng quát, độ chính xác, độ đầy đủ, và độ đầy đủ điều hòa lần lượt là là 87,42%; 86,32%; 85,41%; 85,63% so với 60,34%; 57,73%; 57,75%; 57,16% của bộ từ điển VietSentiWordNet ban đầu.

Bảng 8. Hiệu suất xác định hướng quan điểm giữa từ điển VietSentiWordNetPlus và VietSentiWordNet

TT	Điểm du lịch	N	Pos/ Neg	VSWN				VSWNPlus			
				Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
1	Đại Nội	146	97/41	52,17%	48,60%	48,38%	47,79%	90,58%	89,42%	87,67%	88,47%
2	Núi Bạch Mã`	123	77/42	61,34%	58,06%	58,23%	58,12%	85,71%	85,18%	83,01%	83,89%
3	Suối Thanh Tân	121	84/33	69,23%	62,01%	62,01%	62,01%	84,62%	82,22%	78,25%	79,83%
4	Biển Thuận An	116	62/51	61,06%	60,52%	59,82%	59,67%	85,84%	86,78%	85,01%	85,43%
5	Suối Voi	112	59/51	56,36%	55,80%	55,33%	54,87%	82,73%	84,29%	81,90%	82,20%
6	Chùa Thiên Mụ	109	74/29	54,37%	52,90%	53,56%	51,26%	86,41%	82,93%	84,25%	83,54%
7	Biển Lộc Bình	106	39/42	62,96%	64,41%	63,46%	62,50%	91,36%	91,34%	91,39%	91,35%
8	Biển Cảnh Dương	97	69/23	69,57%	61,96%	63,77%	62,47%	83,70%	78.15%	80.43%	79.14%
9	Đầm Lập An	85	46/34	63,75%	62,83%	62,72%	62,77%	90,00%	90.10%	89.39%	89.68%
10	Chùa Huyền Không	66	37/22	52,54%	50,18%	50,18%	50,12%	93,22%	92.75%	92.75%	92.75%
Trung bình				60,34%	57,73%	57,75%	57,16%	87,42%	86,32%	85,41%	85,63%

Trong đó, VSWN: từ điển VietSentiWordNet, VSWNPlus: từ điển VietSentiWordNetPlus, N: Số câu bình luận, Pos/Neg: tỉ lệ số câu tích cực/ tiêu cực, Accuracy là độ chính xác tổng quát, Precision là độ chính xác, Recall là độ đầy đủ, F-score là độ đầy đủ điều hòa.

5 Kết luận

Trong bài báo này, chúng tôi đã đề xuất một mô hình mở rộng từ điển VietSentiWordNet cho miền dữ liệu thuộc lĩnh vực du lịch sử dụng phương pháp dựa trên từ vựng. Cụ thể, chúng tôi đã mở rộng từ điển VietSentiWordNet với việc làm giàu thêm các từ quan điểm thuộc lĩnh vực du lịch và tích hợp các chức năng tiền xử lý dữ liệu bao gồm thêm dấu, chuẩn hóa láy âm tiết (đối với những từ thể hiện cảm xúc đặc biệt), chuẩn hóa chữ viết tắt, xử lý biểu tượng cảm xúc. Dựa trên kết quả thực nghiệm, từ điển VietSentiWordNetPlus đã cho kết quả phân lớp quan điểm tốt hơn, với trung bình độ chính xác tổng quát, độ chính xác, độ đầy đủ và độ đầy đủ điều hòa lần lượt là 87,42%; 86,32%; 85,41%; 85,63% so với 60,34%; 57,73%; 57,75%; 57,16% của bộ từ điển VietSentiWordNet ban đầu. Tuy nhiên, việc gán trọng số cho các từ quan điểm để xây dựng bộ từ quan điểm tiếng Việt có trọng số (VnOpinionWords) có mức độ chính xác phụ thuộc vào độ chính xác của từ điển Việt – Anh, nên cần có nhiều nghiên cứu hơn nữa để nâng cao hiệu quả của cách tiếp cận khai phá quan điểm dựa trên từ vựng này.

Tài liệu tham khảo

1. A. Arora, C. Patil, S. Correia (2015), *Opinion Mining: An Overview*, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 11, pp. 94-98.
2. W. Medhat, A. Hassan, H. Korashy (2014), *Sentiment analysis algorithms and applications: A survey.* Ain Shams Engineering Journal, 5(4), pp. 1093-1113.
3. C. Bucur (2015), *Using opinion mining techniques in tourism*, in Proceedings of the 2nd Global Conference on Business, Economics, Management and Tourism, Procedia Economics and Finance 23, pp. 1666-1673.
4. S. Baccianella, A. Esuli, F. Sebastiani (2010), *SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining*, In: Proceedings of the 7th Conference on International Language Resources and Evaluation, pp. 2200–2204
5. V. Soni, M. Patel (2014), *Unsupervised Opinion Mining From Text Reviews Using SentiWordNet*, International Journal of Computer Trends and Technology (IJCTT) V11(5), pp. 234-238.
6. F. M. Kundi, A. Khan, S. Ahmad, M. Z. Asghar (2014), *Lexicon-Based Sentiment Analysis in the Social Web*, Journal of Basic and Applied Scientific Research, 4(6), pp. 238-248.
7. G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu, C. Chen (2010), *DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis*. Expert Systems with Applications 37, pp. 6182–6191.
8. Kieu Thanh Binh, Pham Bao Son (2010), *Sentiment Analysis for Vietnamese*, In: 2010 Second International Conference on Knowledge and Systems Engineering, pp. 152–157.
9. Vu Tien Thanh, Pham Huyen Trang, Luu Cong To, Ha Quang Thuy (2011), *A Feature-Based Opinion Mining Model on Product Reviews in Vietnamese*. In Semantic Methods for Knowledge Management and Communication (SCI 381), pp. 23-33.
10. Vu Xuan Son, P. Seong-Bae (2014), *Construction of Vietnamese SentiWordNet by using Vietnamese Dictionary*, The 40th Conference of the Korea Information Processing Society, pp. 745-748, South Korea.
11. P. Haseena Rahmath (2014), *Opinion Mining and Sentiment Analysis challenges and Applications*, International Journal of Application or Innovation in Engineering & Management. Volume 3, Issue 5.
12. Hong Nam Nguyen, Thanh Van Le, Hai Son Le, Tran Vu Pham, (2014). *Domain Specific Sentiment Dictionary for Opinion Mining of Vietnamese Text*. The 8th Multi-Disciplinary International Workshop on Artificial Intelligence (MIWAI 2014), pp. 136-148.
13. A. Sadia, F. Khan, F. Bashir (2018), *An Overview of Lexicon-Based Approach For Sentiment Analysis*, International Electrical Engineering Conference, IEP Centre, Karachi, Pakistan
14. K. Aung, N. Myo (2017), *Sentiment Analysis of Students' Comment Using Lexicon Based Approach*, Computer and Information Science (ICIS), IEEE/ACIS 16th International Conference IEEE, pp. 149-154.
15. B. Liu (2007), *Web Data Mining: Exploring Hyperlinks, Contents and Usage data*, Springer, Second Edition.
16. N. Mishra, C.K.Jha, PhD. (2012), *Classification of Opinion Mining Techniques*, International Journal of Computer Applications, Volume 56 – No.13.
17. Võ Tuyết Ngân, Đỗ Thanh Nghị (2015), *Phân loại ý kiến trên Twitter*, Tạp chí Khoa học Trường Đại học Cần Thơ, pp. 32-38.
18. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser (2017), *Attention Is All You Need*, arXiv:1706.03762v5 [cs.CL].
19. T. Kim, K. Wurster (2015), emoji v.0.3.4, BSD License.
20. Viet Trung Tran (2016), Python Vietnamese Toolkit, MIT License.
21. M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi (2020), *GBSVM: Sentiment Classification from Unstructured Reviews Using Ensemble Classifier*, *Appl. Sci. 10*(8), 2788.

EXPANDING VIETSENTIWORDNET DICTIONARY FOR TOURISM DATA DOMAIN USED A LEXICON-BASED APPROACH

Le Van Hoa*

School of Hospitality and Tourism – Hue University

Abstract. Opinion mining helps to determine the semantic orientations (positive, negative) of customers about a topic, product or service. There are several different approaches to opinion mining, in which the lexicon-based approach to opinion mining is relatively popular. The accuracy of the lexicon-based approach to opinion mining is highly dependent on the dictionary, in which contains opinion expressing words to specific domains. One data set may give very good classification in one domain, but it performs very poor in some other domains. Nowadays, VietSentiWordNet is a Vietnamese opinion dictionary that it is used relatively popular, but it lacks many opinion words for the tourism data domain. This paper focuses on expanding VietSentiWordNet dictionary with enrich opinion words belong to the tourism domain. In which a model for expanding VietSentiWordNet dictionary used a lexicon-based approach is proposed, and process data preprocessing consist of many functions also added to improve the efficiency of opinion classification. Evaluation results show that the expansion of VietSentiWordNet dictionary classifies opinion more accurately for opinion sentences in the tourism domain.

Keywords: VietSentiWordNet dictionary; tourism data domain; opinion mining; lexicon-based approach; semantic orientations.