

BUILDING A RESTAURANT ASSESSMENT SYSTEM IN THUA THIEN HUE PROVINCE BASED ON ONLINE COMMENTS

Le Van Hoa

School of Hospitality and Tourism – Hue University

ARTICLE INFO	ABSTRACT
Received: 22/11/2021 Revised: 10/01/2022 Published: 11/02/2022	Vietnamese opinion mining systems are based on the lexicon-based approach using the VietSentiWordNet dictionary. However, this data dictionary applies to the news domain, so when used to classify in the tourism domain, it will be ineffective and easy to cause confusion. The objective of this paper is to build a restaurant assessment system with high classification efficiency in the tourism domain. To build the system, we use lexicon-based approach to opinion mining combined with the Vietnamese opinion dictionary in the tourism domain VietSentiWordNetPlus. In addition, we also apply data preprocessing techniques to the comments to increase the semantics of the sentences. The experimental results showed that, our system gave better opinion classification results, with average accuracy, precision, recall and F-score 84.64%; 76.39%; 81.12%; 78.15% versus 71.76%; 63.64%; 68.72%; 63.82% of the system uses the VietSentiWordNet dictionary. Our system is highly effective when classifying opinion with data sources in the tourism domain such as restaurants, hotels, tourist attractions.
KEYWORDS	
Opinion mining	
Online comments	
Dictionary	
Data preprocessing	
Tourism domain	

XÂY DỰNG HỆ THỐNG ĐÁNH GIÁ NHÀ HÀNG TRÊN ĐỊA BÀN TỈNH THỪA THIÊN HUẾ DỰA VÀO CÁC BÌNH LUẬN TRỰC TUYẾN

Lê Văn Hòa

Trường Du lịch – ĐH Huế

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 22/11/2021 Ngày hoàn thiện: 10/01/2022 Ngày đăng: 11/02/2022	Các hệ thống khai phá quan điểm tiếng Việt dựa trên phương pháp từ vựng thông thường sử dụng bộ từ điển VietSentiWordNet. Tuy nhiên, bộ từ điển dữ liệu này áp dụng cho miền tin tức nên khi sử dụng để phân lớp trong lĩnh vực du lịch sẽ đạt hiệu quả không cao và dễ gây nhầm lẫn. Mục tiêu của bài báo này nhằm xây dựng hệ thống đánh giá nhà hàng đạt hiệu quả phân lớp cao trong lĩnh vực du lịch. Để xây dựng hệ thống, chúng tôi sử dụng phương pháp khai phá quan điểm dựa trên từ vựng kết hợp với bộ từ điển quan điểm tiếng Việt thuộc lĩnh vực du lịch VietSentiWordNetPlus. Ngoài ra, chúng tôi còn áp dụng các kỹ thuật tiền xử lý dữ liệu cho các câu bình luận để tăng ngữ nghĩa cho câu. Kết quả thực nghiệm cho thấy, hệ thống của chúng tôi 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 84,64%; 76,39%; 81,12%; 78,15% so với 71,76%; 63,64%; 68,72%; 63,82% của hệ thống sử dụng bộ từ điển VietSentiWordNet. Hệ thống của chúng tôi đạt hiệu quả cao khi phân lớp quan điểm với nguồn dữ liệu thuộc lĩnh vực du lịch như: nhà hàng, khách sạn, điểm du lịch.
TỪ KHÓA	
Khai phá quan điểm	
Bình luận trực tuyến	
Từ điển	
Tiền xử lý dữ liệu	
Lĩnh vực du lịch	

DOI: <https://doi.org/10.34238/tnu-jst.5281>

Email: levanhua84@hueuni.edu.vn

<http://jst.tnu.edu.vn>

27

Email: jst@tnu.edu.vn

1. Giới thiệu

Mỗi khách hàng có nhu cầu lựa chọn cho mình một nhà hàng khác nhau tùy theo mục đích tới nhà hàng, sở thích về món ăn, giá cả, không gian và cách phục vụ. Ví dụ, để tổ chức tiệc sinh nhật, khách hàng thường quan tâm đến các nhà hàng có không gian đẹp, bãi đậu xe; hoặc khách hàng thích ăn hải sản thì quan tâm đến các nhà hàng có các món ăn hải sản tươi ngon. Ngày nay, khi lượng đánh giá của khách hàng tăng nhanh trên các trang web đánh giá trực tuyến, điều này vừa mang lại những thuận lợi nhưng cũng tạo ra những thách thức vì khách hàng sẽ mất nhiều thời gian để tìm kiếm và thu thập thông tin hữu ích theo các đặc trưng khác nhau của nhà hàng từ rất nhiều đánh giá trực tuyến nhằm đưa ra quyết định lựa chọn nhà hàng. Ngoài ra, thông tin đánh giá về nhà hàng có thể bị sai lệch nếu chỉ phân tích một số đánh giá hoặc chỉ phân tích đánh giá ở duy nhất một nguồn dữ liệu. Hiện nay, các hệ thống đánh giá, tư vấn trong các website nhà hàng chỉ nhằm mục đích đánh giá và so sánh giữa các nhà hàng hay món ăn dựa vào điểm số đánh giá hoặc dựa vào việc xếp hạng có gắn sao. Bởi vì, các website này chưa quan tâm đến việc đánh giá và so sánh dựa vào các bình luận trực tuyến của khách hàng. Trong khi đó, các bình luận trực tuyến là một trong những thông tin có độ tin cậy cao và ảnh hưởng rất lớn đến quyết định lựa chọn nhà hàng của khách hàng. Khi mọi người có ý định chọn nhà hàng, họ sẽ kiểm tra các đánh giá hoặc xếp hạng của các nhà hàng đó trên các trang web trực tuyến như Foody.vn, Tripadvisor.com.vn,... trước khi chọn chúng. Mọi người sẽ chọn nhà hàng dựa trên những cảm nhận tích cực trong các đánh giá về nó [1].

Với sự bùng nổ của dữ liệu lớn (big data) và công nghệ Internet kết nối vạn vật (Internet of Things), các ý kiến đánh giá trực tuyến của khách hàng cần được thu thập, khai thác và tổng hợp một cách tự động bằng các hệ thống máy tính, cho phép các nhà kinh doanh có thể dễ dàng theo dõi hành vi mua sắm, phát hiện sở thích và đánh giá sự hài lòng của khách hàng về chất lượng sản phẩm, dịch vụ [2]. Đồng thời, khách hàng cũng cần thông tin tổng hợp ý kiến đánh giá của cộng đồng để có những quyết định mua sắm của mình. Chính vì thế, khai thác quan điểm tự động đã trở thành tiêu điểm của rất nhiều nghiên cứu trong các lĩnh vực khác nhau [3]. Trong những năm gần đây, khai phá quan điểm dựa trên từ vựng là một hướng nghiên cứu đang được nhiều nhà khoa học quan tâm [4]-[6]. Trong đó, nghiên cứu [4] đã sử dụng từ điển VietSentiWordNet để xây dựng hệ thống đánh giá điểm du lịch trên địa bàn tỉnh Thừa Thiên Huế dựa vào bình luận của người dùng facebook. Tuy nhiên, do chính sách của facebook nên tác giả sử dụng nguồn dữ liệu là các fanpage do chính tác giả xây dựng, do đó độ tin cậy về dữ liệu thu thập chưa cao. Một nghiên cứu khác của Cristian [5] đã xây dựng hệ thống khai phá quan điểm để trích xuất các đánh giá từ Internet và phân loại chúng dựa vào từ điển SentiWordNet. Ngoài ra, Vibha và cộng sự [6] sử dụng phương pháp từ vựng dựa vào từ điển SentiWordNet để tìm ra khía cạnh tích cực và tiêu cực của sản phẩm điện thoại di động trên website Amazon.com.

Trong bài báo này, chúng tôi tập trung vào việc xây dựng hệ thống khai phá quan điểm dựa vào phương pháp từ vựng áp dụng cho miền dữ liệu nhà hàng với nguồn dữ liệu là các bình luận trực tuyến chủ yếu trên hai trang Foody.vn và Tripadvisor.com.vn. Trong quá trình thiết kế mô hình hệ thống, chúng tôi đã sử dụng phương pháp từ vựng kết hợp với bộ từ điển quan điểm tiếng Việt VietSentiWordNetPlus [7] được mở rộng từ bộ từ điển VietSentiWordNet của Vũ Xuân Sơn và cộng sự [8] với nhiều bổ sung liên quan đến các từ thể hiện quan điểm, cảm xúc thuộc lĩnh vực du lịch. Ngoài ra, trong mô hình này, chúng tôi cũng đã sử dụng các kỹ thuật tiền xử lý dữ liệu nhằm xây dựng một hệ thống khai phá quan điểm thực hiện việc phân lớp quan điểm đạt hiệu quả cao.

2. Nghiên cứu liên quan

Đã có một số nghiên cứu liên quan đến hệ thống khai phá quan điểm trong lĩnh vực nhà hàng. Cụ thể, nghiên cứu [9] cho rằng, đánh giá của khách hàng về nhà hàng đóng một vai trò quan trọng trong quá trình ra quyết định. Khi khách hàng quyết định một nhà hàng, khía cạnh quan

trọng nhất mà họ xem xét là loại thức ăn mà nhà hàng phục vụ, chất lượng của món ăn. Ngoài ra, nhóm tác giả đã phát triển một quy trình tổng thể về xếp hạng nhà hàng dựa vào khai phá quan điểm bằng cách sử dụng thuật toán cây quyết định. Tuy nhiên, nhóm tác giả chỉ quan tâm đến dữ liệu xếp hạng nhà hàng nhưng chưa quan tâm đến các bình luận tích cực, tiêu cực theo từng khía cạnh. Ngoài ra, nghiên cứu này dựa trên một nguồn dữ liệu được trích xuất từ tập dữ liệu xếp hạng nhà hàng Kaggle nên hạn chế về dữ liệu nghiên cứu. Trong khi đó, nghiên cứu [1] đã thực hiện việc khai phá quan điểm dựa trên khía cạnh sử dụng các đánh giá trực tuyến của khách hàng về các nhà hàng ở Indonesia. Các khía cạnh được phân loại là tích cực nếu đánh giá đề cập đến các cụm từ tích cực như: ngon, sạch, rẻ và xuất sắc. Các khía cạnh được phân loại là tiêu cực nếu đánh giá đề cập đến các cụm từ tiêu cực như: xấu, đắt, bẩn và chậm. Hệ thống dựa vào các bình luận về nhà hàng để phân các câu quan điểm thành 3 lớp (tích cực, tiêu cực, trung lập) theo các khía cạnh (món ăn, giá cả, dịch vụ và môi trường xung quanh,...). Tuy nhiên, hệ thống sử dụng tập dữ liệu với các ngôn ngữ trộn lẫn, điều này dễ gây nhầm lẫn cho mô hình phân lớp quan điểm. Ngoài ra, nghiên cứu [10] đã đề xuất một hệ thống để so sánh các sản phẩm, thực hiện các khuyến nghị cho khách hàng và đưa ra kết quả trực quan. Mọi người có thể so sánh các sản phẩm ở cấp độ tính năng để giúp khách hàng đưa ra quyết định sáng suốt. Hơn nữa, khách hàng có thể thấy rõ điểm mạnh và điểm yếu của từng sản phẩm thông qua so sánh. Tuy nhiên, bài báo chỉ quan tâm đến xếp hạng theo từng đặc trưng của sản phẩm mà chưa quan tâm đến yếu tố tích cực, tiêu cực và các câu bình luận liên quan đến các sản phẩm.

Ở trong nước, nghiên cứu [11] đề xuất phương pháp khai thác ý kiến và phân tích cảm xúc khách hàng thông qua việc thu thập tập dữ liệu là ý kiến bình luận của khách hàng trên website Foody.vn - một trang thương mại điện tử hàng đầu trong lĩnh vực dịch vụ đặt hàng trực tuyến. Nhóm tác giả đã tiến hành thực nghiệm bằng phương pháp học máy để khai phá ý kiến từ bình luận dạng văn bản của khách hàng và trực quan hóa kết quả hỗ trợ ra quyết định. Kết quả thực nghiệm cho thấy độ chính xác 90% của phương pháp đề xuất và kết quả khai phá được tập thông tin, tri thức tiềm ẩn có giá trị từ tập ngữ liệu nhằm giúp các cửa hàng, nhà quản trị hiểu được các ưu nhược điểm về sản phẩm, dịch vụ để cải thiện chiến lược kinh doanh tốt hơn. Tuy nhiên, nhóm tác giả chưa xử lý biểu tượng cảm xúc, đây là một trong những yếu tố có thể quyết định khả năng phân loại quan điểm của hệ thống. Một hạn chế khác, nhóm tác giả chỉ thu thập dữ liệu từ website Foody.vn nên bị giới hạn về dữ liệu nghiên cứu. Ngoài ra, nghiên cứu [12] trình bày một phương pháp phân tích quan điểm người dùng dựa trên các nhận xét cá nhân. Bài báo này tập trung vào giải quyết ba nhiệm vụ của bài toán phân tích quan điểm: nhận dạng và trích rút nội dung theo từng khía cạnh; khám phá việc người dùng xếp hạng trên từng khía cạnh đối với sản phẩm; dự đoán trọng số xếp hạng của các khía cạnh trong mỗi nhận xét. Kết quả thực nghiệm trên ba bộ dữ liệu cà phê, bia, khách sạn cho thấy độ chính xác của phương pháp đề xuất là khá tốt cho cả bài toán trích rút khía cạnh cũng như cho bài toán dự đoán xếp hạng khía cạnh. Tuy nhiên, nhóm tác giả chưa quan tâm đến các nhận xét tích cực, tiêu cực mà chỉ quan tâm đến trọng số xếp hạng của các khía cạnh.

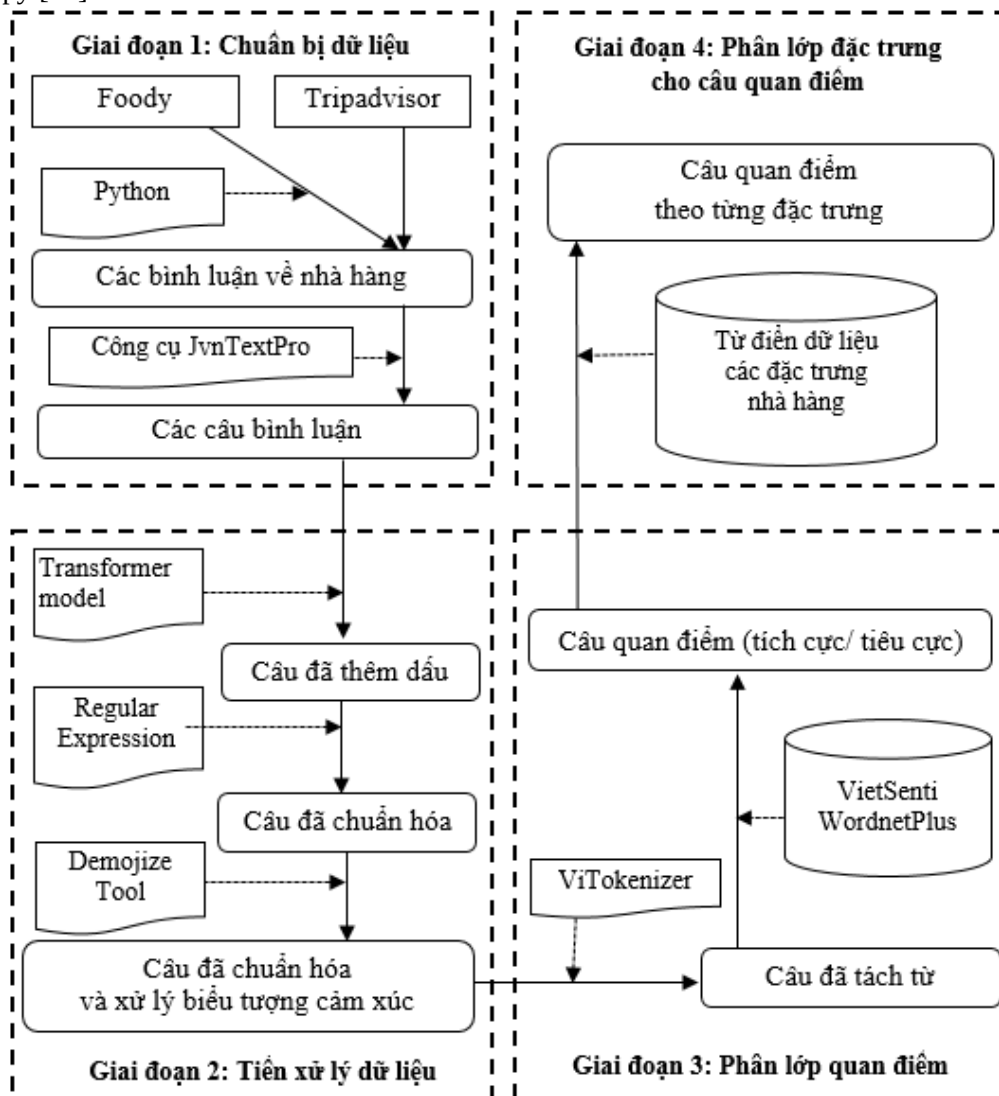
3. Mô hình hệ thống khai phá quan điểm dựa vào phương pháp từ vựng áp dụng cho miền dữ liệu thuộc lĩnh vực nhà hàng

Hình 1 mô tả mô hình của hệ thống khai phá quan điểm dựa vào phương pháp từ vựng áp dụng cho miền dữ liệu thuộc lĩnh vực nhà hàng. Mô hình bao gồm 4 giai đoạn thực hiện như sau: (1) Chuẩn bị dữ liệu (2) Tiền xử lý dữ liệu (3) Phân lớp quan điểm (4) Phân lớp đặc trưng cho câu quan điểm.

3.1. Giai đoạn 1: Chuẩn bị dữ liệu

Để thu thập dữ liệu từ các trang đánh giá trực tuyến, chúng tôi sử dụng bộ thư viện Python. Bộ thư viện này cho phép thu thập các đánh giá trực tuyến theo từng nhà hàng. Sau khi đã thu thập được các bình luận về nhà hàng, chúng tôi dựa vào công cụ JvnTextPro để thực hiện tách câu đối

với những bình luận có nhiều hơn 2 câu. Công cụ JvnTextPro được sử dụng rất hiệu quả để xử lý văn bản tiếng Việt trên nền tảng Java với thuật toán Conditional Random Fields và Maximum Entropy [13].



Hình 1. Mô hình hệ thống khai phá quan điểm sử dụng phương pháp dựa vào từ vựng

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 tăng ngữ nghĩa cho các câu bình luận, chúng tôi tiến hành thêm dấu cho câu đối với các câu tiếng Việt không dấu. Bài toán 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 [14]. Trong giai đoạn này, chúng tôi còn tiến hành chuẩn hóa dữ liệu tiếng Việt sử dụng các kỹ thuật 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 “Hải sản ngonnn quá điiiiiiii!!!!!!” sẽ được chuẩn hóa thành “Hải sản ngon quá đi!” hoặc “Món ăn quá tuyệt vờiiiiiii” sẽ được chuẩn hóa thành “Món ăn quá tuyệt vời”. 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 dựa vào công cụ Demojize trong ngôn ngữ lập trình Python để xử lý biểu tượng cảm xúc bằng cách chuyển các biểu tượng cảm xúc này thành văn bản.

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, chuẩn hóa. Chúng tôi dựa vào công cụ ViTokenizer để thực hiện tách từ trong câu. Công cụ ViTokenizer sử dụng thuật toán Conditional Random Field với độ chính xác tách từ tiếng Việt hơn 97,86%. Công việc tiếp theo của giai đoạn này là xác định hướng quan điểm của câu, chúng tôi sử dụng phương pháp từ vựng kết hợp với từ điển VietSentiWordnetPlus thuộc lĩnh vực du lịch. Từ điển VietSentiWordNetPlus [7] được mở rộng từ bộ từ điển VietSentiWordNet của Vũ Xuân Sơn và cộng sự [8] với việc bổ sung hơn 1.710 từ thể hiện quan điểm, cảm xúc thuộc lĩnh vực du lịch. Từ điển VietSentiWordNet của Vũ Xuân Sơn áp dụng cho miền tin tức nên khi áp dụng vào lĩnh vực du lịch để phân lớp quan điểm sẽ gây ra hiểu nhầm cũng như không phát hiện ra một số từ quan điểm thuộc lĩnh vực du lịch, dẫn đến kết quả phân lớp không chính xác. Trong mô hình này, chúng tôi sử dụng từ điển VietSentiWordnetPlus nên đã cải thiện được khả năng phân lớp quan điểm của hệ thống.

3.4. Giai đoạn 4: Phân lớp đặc trưng cho câu quan điểm

Để thực hiện việc phân lớp đặc trưng cho các câu quan điểm, chúng tôi xây dựng bộ từ điển các đặc trưng về nhà hàng. Các đặc trưng này chủ yếu được tham khảo từ nghiên cứu của Nurifan và cộng sự [15], cấu trúc và nội dung bộ từ điển các đặc trưng nhà hàng được minh họa như trong Bảng 1. Dựa vào bộ từ điển các đặc trưng nhà hàng, chúng tôi có thể phân lớp các câu quan điểm vào từng đặc trưng của nhà hàng.

Bảng 1. Cấu trúc và nội dung bộ từ điển các đặc trưng nhà hàng

TT	Mã đặc trưng	Tên đặc trưng	Các thuộc tính đi kèm
1	Nha_hang	Nhà hàng	Thuộc tính chung, chất lượng, giá cả, kiểu
2	Khong_gian	Không gian	Thuộc tính chung
3	Vi_tri	Vị trí	Thuộc tính chung
4	Do_an	Đồ ăn	Chất lượng, giá cả, kiểu
5	Phuc_vu	Phục vụ	Thuộc tính chung, chất lượng
6	Do_uong	Đồ uống	Chất lượng, giá cả, kiểu

4. Thử nghiệm và phân tích kết quả

Trong thử 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 [16]. 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 2.

Bảng 2. 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):

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

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):

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (2)$$

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):

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (3)$$

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):

$$\text{F-score} = 2 \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \quad (4)$$

Một ví dụ minh họa về kết quả đánh giá thực nghiệm của hệ thống phân lớp sử dụng từ điển VietSentiWordNetPlus đối với nhà hàng Bà Đồ được mô tả như trong Bảng 3.

Bảng 3. Kết quả đánh giá thực nghiệm đối với nhà hàng Bà Đồ

		Mẫu dữ liệu thực tế		Kết quả đánh giá			
		Positive	Negative	Accuracy	Precision	Recall	F-score
Hệ thống phân lớp	Positive	TP là 38	FP là 02	84,13%	95,00%	82,61%	88,37%
	Negative	FN là 08	TN là 15		65,22%	88,24%	75,00%
Trung bình					80,11%	85,42%	81,69%

Từ kết quả đánh giá thực nghiệm trong Bảng 3, chúng ta thấy hiệu suất phân lớp sử dụng từ điển VietSentiWordNetPlus đối với nhà hàng Bà Đồ là khá cao. Trong đó, hệ thống dự đoán một lớp là Positive có Precision(Positive) là 95,00%, Recall(Positive) là 82,61%, F-score(Positive) là 88,37%; dự đoán một lớp là Negative có Precision(Negative) là 65,22%, Recall(Negative) là 88,24%, F-score(Negative) là 75,00%; độ chính xác tổng quát (Accuracy) là 84,13%. Như vậy, hiệu suất phân lớp trung bình sử dụng từ điển VietSentiWordNetPlus đối với nhà hàng Bà Đồ có độ chính xác tổng quát (Accuracy) là 84,13%, độ chính xác (Precision) là 80,11%, độ đầy đủ (Recall) là 85,42% và độ đầy đủ điều hòa (F-score) là 81,69%.

Quá trình thực nghiệm hệ thống đánh giá nhà hàng trên địa bàn tỉnh Thừa Thiên Huế dựa vào các bình luận trực tuyến bao gồm: (1) xác định nguồn dữ liệu - nguồn dữ liệu cung cấp cho hệ thống là các bình luận trực tuyến được chúng tôi thu thập từ hai trang Foody.vn và Tripadvisor.com.vn. (2) xác định số lượng dữ liệu - chúng tôi chọn ra 10 nhà hàng nhận được nhiều bình luận nhất từ phía khách hàng trên hai trang Foody.vn và Tripadvisor.com.vn. Sau khi lọc bỏ các bình luận không phù hợp, dữ liệu còn lại để tiến hành thực nghiệm là hơn 760 câu bình luận của 10 nhà hàng trên địa bàn tỉnh Thừa Thiên Huế (3) Các giai đoạn xử lý - chúng tôi tiến hành thực nghiệm theo các giai đoạn như trong mục số 3 của bài báo (mô hình hệ thống khai phá quan điểm dựa vào phương pháp từ vựng áp dụng cho miền dữ liệu thuộc lĩnh vực nhà hàng). Trong các giai đoạn này, để tăng ngữ nghĩa cho các câu bình luận, chúng tôi áp dụng các kỹ thuật tiền xử lý dữ liệu như: thêm dấu, chuẩn hóa lấy âm tiết, chuẩn hóa chữ viết tắt, xử lý biểu tượng

cảm xúc; để xác định hướng quan điểm của câu, chúng tôi sử dụng phương pháp từ vựng kết hợp với từ điển VietSentiWordnetPlus thuộc lĩnh vực du lịch; và để phân lớp các câu quan điểm vào từng đặc trưng của nhà hàng, chúng tôi dựa vào bộ từ điển các đặc trưng nhà hàng.

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. Bảng 4 so sánh hiệu suất xác định hướng quan điểm (theo phương pháp Accuracy và Precision - Recall) của 10 nhà hàng 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 84,64%; 76,39%; 81,12%; 78,15% so với 71,76%; 63,64%; 68,72%; 63,82% của bộ từ điển VietSentiWordNet.

Bảng 4. Hiệu suất xác định hướng quan điểm giữa từ điển VietSentiWordNetPlus và VietSentiWordNet

TT	Nhà hàng	N	Pos/ Neg	VSWN				VSWNPlus			
				Accuracy (%)	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
1	Bà Đỏ	69	46/17	80,95	76,19	79,54	77,39	84,13	80,11	85,42	81,69
2	Cơm Bà Thu	66	50/9	76,27	62,12	67,78	63,39	86,44	74,15	78,33	75,92
3	DMZ Bar	48	27/14	70,73	68,13	69,18	68,46	80,49	78,38	80,03	78,97
4	Golden Rice	59	47/6	69,81	60,98	75,71	59,23	90,57	76,64	87,41	80,59
5	Hot Tuna	65	43/11	74,07	66,77	73,57	67,69	87,04	79,64	85,10	81,78
6	Nook Eatery	63	45/11	66,07	55,99	58,28	55,80	83,93	74,62	76,26	75,38
7	Phố Cổ	61	42/14	69,64	64,29	67,86	64,68	82,14	76,61	80,95	78,13
8	Quán Hạnh	97	73/15	75,00	64,42	71,69	65,60	85,23	74,21	77,85	75,76
9	Spice Viet	158	92/39	67,94	62,87	63,88	63,20	78,38	76,37	77,46	76,79
10	Zucca	75	59/8	67,16	54,65	59,75	52,76	88,06	73,18	82,42	76,49
Trung bình				71,76	63,64	68,72	63,82	84,64	76,39	81,12	78,15

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 xây dựng hệ thống đánh giá nhà hàng sử dụng phương pháp khai phá quan điểm dựa trên từ vựng. Hệ thống chúng tôi xây dựng đạt hiệu quả phân lớp câu quan điểm cao nhờ việc sử dụng từ điển quan điểm tiếng Việt VietSentiWordNetPlus và áp dụng các kỹ thuật tiền xử lý dữ liệu bao gồm các kỹ thuật: thêm dấu, chuẩn hóa lấy âm tiết, chuẩn hóa chữ viết tắt, xử lý biểu tượng cảm xúc. Dựa trên kết quả thực nghiệm, hệ thống sử dụng từ điển VietSentiWordNetPlus đã cho kết quả phân lớp quan điểm tốt hơn so với hệ thống sử dụng bộ từ điển VietSentiWordNet. Trong thời gian tới, chúng tôi tiếp tục tìm hiểu thêm các kỹ thuật thu thập dữ liệu đối với nguồn dữ liệu trên các fanpage facebook, diễn đàn để nâng cao hiệu quả của hệ thống.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] S. Andi and B. Indra, "Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia," In *International Conference on Asian Language Processing (IALP)*, Shanghai, 2019, pp. 59-64.
- [2] C. C. Jovelyn, L. Jesserine, and J. R. Danica, "Text mining customer reviews for aspectbased restaurant rating," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 10, no. 6, pp. 43-52, 2018.
- [3] P. K. Thai, T. A. Nguyen, and H. T. T. Tran, "A support system for tourism services assessment and recommendation based on pinion mining online customer reviews," *Journal of Science and Technology - Industrial University of HCMC*, vol. 46, pp. 175-189, 2020.

-
- [4] H. V. Le, "Building a opinion mining model based on comments by facebook users: the case of tourist attractions in Thua Thien Hue," *Hue University Journal of Science: Techniques and Technology*, vol. 128, no. 2A, pp. 69-81, 2018.
 - [5] B. Cristian, "Using opinion mining techniques in tourism," *Procedia Economics and Finance*, vol. 23, no. 8, pp. 1666-1673, 2015.
 - [6] S. Vibha and P. Meenakshi, "Unsupervised Opinion Mining From Text Reviews Using SentiWordNet," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 11, no. 5, pp. 234-238, 2014.
 - [7] H. V. Le, "Expanding Vietsentiwordnet dictionary for tourism data domain used a lexicon-based approach," *Hue University Journal of Science: Techniques and Technology*, vol. 129, no. 2A, pp. 15-29, 2020.
 - [8] S. X. Vu and P. Seong-Bae, "Construction of Vietnamese SentiWordNet by using Vietnamese Dictionary," In The 40th Conference of the Korea Information Processing Society, South Korea, 2014.
 - [9] M. Kavin, S. Aravinth, D. Hari, and M. Monica, "Opinion Mining on Restaurant Rating Based on Aspects," *International Journal of Computational Science and Engineering*, vol. 10, no. 1, pp. 25-34, 2020.
 - [10] S. Jianshu, L. Chong, Z. Xiaoyan, and H. Minlie, "Mining Reviews for Product Comparison and Recommendation," *Polibits*, vol. 39, no. 39, pp. 33-40, 2009.
 - [11] B. D. L. Nguyen, H. V. Nguyen, and T. T. Ho, "A text-based model for opinion mining and sentiment analysis from online customer reviews in food industry," *Ho Chi Minh City Open University Journal of Science*, vol. 16, no. 1, pp. 64-78, 2020.
 - [12] T. T. N. Nguyen, L. D. Nguyen, G. K. Nguyen, H. T. T. Nguyen, and A. V. Nguyen, "A method of opinion mining by consumers for product quality based on reviews," in *Proceedings of the 10th national conference on fundamental and applied information technology research*, Da Nang, 2017, pp. 585-594.
 - [13] T. C. Nguyen, H. X. Phan, and T. T. Nguyen, "JVnTextPro: A Java-based Vietnamese Text Processing Tool," 2010. [Online]. Available: <http://jvntextpro.sourceforge.net/>. [Accessed Jul. 15, 2021].
 - [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and L. Kaiser, "Attention Is All You Need," *Proceedings of the 31st International Conference on Neural Information Processing System*, vol. 5, no. 11, pp. 6000-6010, 2017.
 - [15] N. Farza, S. Riyanarto, and S. R. Kelly, "Aspect Based Sentiment Analysis for Restaurant Reviews Using Hybrid ELMoWikipedia and Hybrid Expanded Opinion Lexicon-SentiCircle," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 6, pp. 47-58, 2019.
 - [16] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, and G. S. Choi, "GBSVM: Sentiment Classification from Unstructured Reviews Using Ensemble Classifier," *Applied Sciences*, vol. 10, no. 8, pp. 2788-2807, 2020.