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

Tập 132, Số 2A, 2023, Tr. 63–73; DOI: 10.26459/hueunijtt.v132i2A.7250

XÂY DỰNG MÔ HÌNH PHÂN LỚP BÌNH LUẬN VỀ NHÀ HÀNG DỰA VÀO PHƯƠNG PHÁP HỌC MÁY

Lê Văn Hòa[1]*, Đào Thị Minh Trang[1], Sử Minh Đạt[2], Nguyễn Dương Thiện[1]

[1] Trường Du lịch – Đại học Huế, Việt Nam
[2] Trường Cao Đẳng Công Nghiệp Huế, Việt Nam

Tóm tắt. Học máy được biết đến là phương pháp phân lớp hiệu quả, trong đó hiệu quả của việc phân lớp khá phụ thuộc vào các đặc tính của đối tượng được trích chọn từ bộ dữ liệu. Với thực tế bùng nổ dữ liệu như hiện nay, đặc biệt là dữ liệu bình luận trên các website đánh giá trực tuyến, việc phân lớp gặp phải không ít thách thức. Tuy nhiên, việc phân lớp chính xác sẽ mang lại những ý nghĩa to lớn cho hoạt động tư vấn. Bài báo này nghiên cứu và xây dựng một mô hình phân lớp dựa trên học máy đối với dữ liệu là các bình luận về nhà hàng. Dữ liệu được thu thập từ các website đánh giá trực tuyến như Tripadvisor.com.vn và Foody.vn. Một kỹ thuật tiền xử lý các câu bình luận nhằm tăng ngữ nghĩa cũng được đề xuất. Thực nghiệm sử dụng 4 thuật toán học máy: Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT) và K-Nearest Neighbor (KNN). Kết quả cho thấy SVM cho kết quả phân lớp tốt nhất khi so sánh với NB, DT và KNN.

Từ khóa: Thuật toán học máy, bình luận, dữ liệu về nhà hàng

Building a restaurant comment classification model based on machine learning methods

Le Van Hoa[1]*, Dao Thi Minh Trang[1], Su Minh Dat[2], Nguyen Duong Thien[1]

[1] School of Hospitality and Tourism - Hue University
[2] Hue Industrial College, Vietnam

Abstract. Machine learning has been recognized as an efficient classification method, in which the effectiveness of the classification significantly depends on the characteristics of the object extracted from the dataset. With the current explosion of data in general, and especially comment data on online review websites in particular, classification faces numerous challenges. However, accurate classification will have significant implications for consulting activities. This study aimed to build a classification model based on machine learning for restaurant comments data. Data is collected from online review websites such as Tripadvisor.com.vn and Foody.vn. A technique for preprocessing comments to enhance semantic understanding is also proposed. Experiments were conducted using four machine

** Liên hệ: levanhoa84@hueuni.edu.vn*

Ngày nhận bài: 17–07–2023; Ngày nhận đăng: 10–08–2023

learning algorithms: Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT), and K-Nearest Neighbor (KNN). The results demonstrate that SVM achieves the best classification outcome in comparison to NB, DT, and KNN.

Keywords: Machine learning algorithm, comment, restaurant data.

1 Giới thiệu

Sự phát triển của Internet làm cho thông tin lưu trữ trực tuyến hàng ngày gia tăng nhanh chóng. Do vậy, để tìm đúng thông tin mà chúng ta cần quan tâm thì mất khá nhiều thời gian nên cần phải dùng những kỹ thuật tổ chức và xử lý dữ liệu về văn bản. Kỹ thuật này được gọi là phân lớp văn bản hay nói cách khác là phân loại văn bản [1]. Phân lớp văn bản là một trong những nhiệm vụ cơ bản của xử lý ngôn ngữ tự nhiên, được ứng dụng rộng rãi trong phân tích quan điểm, phát hiện spam, gắn nhãn chủ đề, phát hiện ý định… Phân lớp văn bản là một kỹ thuật máy học tự động gán các nhãn cho văn bản. Sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên và máy học, bộ phân lớp văn bản có thể phân tích và sắp xếp văn bản theo danh mục, chủ đề và ý định của khách hàng. Với sự bùng nổ của các nguồn thông tin trên Web, mạng xã hội, website đánh giá trực tuyến … làm cho lĩnh vực phân lớp văn bản ngày càng trở nên quan trọng và thu hút nhiều nhà nghiên cứu [2].

Ngày nay, việc viết đánh giá về một sản phẩm hoặc dịch vụ là điều phổ biến. Theo nghiên cứu trong [3], 84% khách hàng tin tưởng vào các đánh giá trực tuyến giống như lời tư vấn hay giới thiệu của bạn bè, người thân. Do đó, các website đánh giá trực tuyến đã trở nên có ích đối với khách hàng. Doanh nghiệp xem những đánh giá đó như là phản hồi cho sản phẩm, dịch vụ của mình. Dữ liệu phản hồi cho phép doanh nghiệp phân tích điểm mạnh và điểm yếu để cải thiện dịch vụ hoặc sản phẩm.

Trong phân lớp văn bản, dữ liệu là một trong những yếu tố quan trọng đóng vai trò quyết định đến kết quả phân lớp văn bản có độ chính xác cao. Đặc thù của nguồn dữ liệu lấy từ các website đánh giá trực tuyến là các bình luận của khách hàng không theo một chuẩn nhất định, người đăng bình luận thường sử dụng những biểu tượng cảm xúc, tiếng lóng và tiếng Việt không dấu. Do đó, để có nguồn dữ liệu đưa vào phân tích đúng chuẩn là một trong những nhiệm vụ quan trọng của bài toán phân lớp văn bản. Ngoài ra, việc chọn đúng phương pháp phân lớp văn bản đối với lĩnh vực nhà hàng là vấn đề cần quan tâm. Với đặc thù dữ liệu thu thập trên các website đánh giá trực tuyến chỉ ở mức độ sơ cấp và lượng dữ liệu rất lớn, các doanh nghiệp không thể dựa vào dữ liệu thô này để ra quyết định. Họ cần biết các tri thức được phân tích từ tập dữ liệu này.

Để giải quyết bài toán này, nghiên cứu của chúng tôi đề xuất phương pháp học máy áp dụng việc phân lớp theo các đặc tính của nhà hàng. Nguồn dữ liệu đưa vào phân lớp là các bình

luận của khách hàng về nhà hàng được thu thập từ các website đánh giá trực tuyến như Tripadvisor.com.vn và Foody.vn. Kết quả nghiên cứu nhằm giúp các nhà quản lý doanh nghiệp nhà hàng nắm bắt thông tin một các dễ dàng và nhanh chóng, từ đó việc kinh doanh được cải thiện, nâng cao sự hài lòng của khách hàng và giữ chân khách hàng tốt hơn.

2 Một số nghiên cứu liên quan

Đã có một số nghiên cứu ngoài nước liên quan đến hệ thống khai phá quan điểm trong lĩnh vực nhà hàng. Cụ thể, nghiên cứu trong [4] 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 chọn một nhà hàng, đặc tính 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 cây quyết định. Tuy nhiên, nhóm tác giả chỉ quan tâm đến dữ liệu xếp hạng nhà hàng mà chưa quan tâm đến các bình luận theo từng đặc tính nhà hàng. 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 phân tích.

Trong [5], Suciati và cộng sự đã thực hiện việc khai phá quan điểm dựa trên đặc tính sử dụng các đánh giá trực tuyến của khách hàng về các nhà hàng ở Indonesia. Các đặc tính được phân lớp là tích cực nếu đánh giá đề cập đến các cụm từ tích cực như: ngon, sạch, rẻ và xuất sắc. Các đặc tính được phân lớp 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 đặc tính (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.

Với các nghiên cứu trong nước về khai phá dữ liệu thuộc lĩnh vực khách sạn - nhà hàng, nghiên cứu trong [6] đã đề xuất mô hình kiến trúc hệ thống cùng với các giải pháp hỗ trợ đánh giá và khuyến nghị dịch vụ du lịch dựa trên khai phá quan điểm. Dữ liệu thực nghiệm nghiên cứu là những bình luận của du khách về các khách sạn tại các tỉnh và thành phố lớn tại Việt Nam. Dữ liệu được thu thập tự động trên trang web Agoda. Trên cơ sở các kết quả thực nghiệm, nghiên cứu đưa ra một số khuyến nghị để có thể triển khai hệ thống này trong thực tiễn ngành du lịch. Nghiên cứu này có giá trị tham chiếu cho các nhà nghiên cứu không chỉ trong lĩnh vực du lịch mà còn trong các lĩnh vực kinh doanh và quản lý. Tuy nhiên, nghiên cứu này vẫn còn nhiều hạn chế như: Nghiên cứu này chỉ thu thập dữ liệu là các bình luận của khách hàng về khách sạn trên trang web Agoda nên hạn chế về đối tượng, phạm vi và dữ liệu phân tích.

Nghiên cứu trong [7] đề xuất một phương pháp phân tích quan điểm người sử dụng dựa trên các nhận xét cá nhân. Nghiên cứu tập trung vào giải quyết ba nhiệm vụ của bài toán khai phá quan điểm: Nhận dạng và trích rút nội dung theo từng đặc tính; Khám phá việc người dùng xếp hạng trên từng đặc tính đối với sản phẩm; và dự đoán trọng số xếp hạng của các đặc tính 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 đặc tính cũng như cho bài toán dự đoán xếp hạng đặc tính. 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 đặc tính.

Một nghiên cứu khác trong [8] đề xuất phương pháp khai phá quan điểm 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 khá cao 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 lớp 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 phân tích.

Dựa vào thực trạng của các nghiên cứu trên, chúng tôi nhận thấy rằng:

- Chủ yếu dữ liệu xếp hạng nhà hàng được quan tâm, trong khi các bình luận theo từng đặc tính nhà hàng chưa được chú ý đến;

- Chưa quan tâm đến các nhận xét tích cực, tiêu cực mà chỉ xem xét trọng số xếp hạng của các đặc tính;

- Chưa xử lý các biểu tượng cảm xúc; đây là một trong những yếu tố có thể quyết định khả năng phân lớp của hệ thống; và

- Chỉ thu thập dữ liệu là các bình luận của khách hàng trên duy nhất một nguồn nên hạn chế về đối tượng, phạm vi và dữ liệu phân tích.

Mô hình được đề xuất trong bài báo sẽ khắc phục được các hạn chế này.

3 Những thách thức đối với bài toán khai phá quan điểm

Với đặc thù nguồn dữ liệu lấy từ các bình luận trên các trang đánh giá trực tuyến và ngôn ngữ sử dụng tiếng Việt, bài toán khai phá quan điểm có thể gặp nhiều khó khăn. Đối với

các bình luận, tùy thuộc vào trình độ kiến thức, nghề nghiệp và tuổi tác mà khách hàng có cách hành văn khác nhau. Các thách thức trong khai phá quan điểm bao gồm:

- Một từ quan điểm có thể có các ý nghĩa trái ngược nhau: tích cực trong tình huống này nhưng tiêu cực trong tình huống khác [9]. Ví dụ, nếu bình luận "Thời gian lên món ăn rất nhanh chóng" thì từ "nhanh chóng" có hàm ý tích cực, nhưng nếu bình luận "Chúng tôi rời nhà hàng rất nhanh chóng", thì từ "nhanh chóng" có hàm ý tiêu cực.

- Một từ quan điểm phụ thuộc vào lĩnh vực [10]. Các đặc tính của một từ quan điểm chỉ có thể phù hợp trong một lĩnh vực nhất định. Ví dụ, trong lĩnh vực nhà hàng nếu bình luận "Thiết kế nhà hàng với kiến trúc cổ xưa" thì từ "cổ xưa" có hàm ý tích cực, nhưng trong lĩnh vực thiết bị điện tử nếu bình luận "Máy tính thời cổ xưa", thì từ "cổ xưa" có hàm ý tiêu cực.

- Xử lý đặc tính ẩn của đối tượng. Ví dụ, trong câu "Món ăn này khá đắt" thì đặc tính "Mức giá" của món ăn bị ẩn.

- Lỗi chính tả, sử dụng sai ngữ pháp, dấu câu bị thiếu, văn bản tiếng Việt không có dấu, chữ viết tắt và sử dụng tiếng lóng. Ví dụ, trong câu "Chai rượu này giá 5 củ thì đắt quá" thì từ "củ" là từ tiếng lóng trong tiếng Việt.

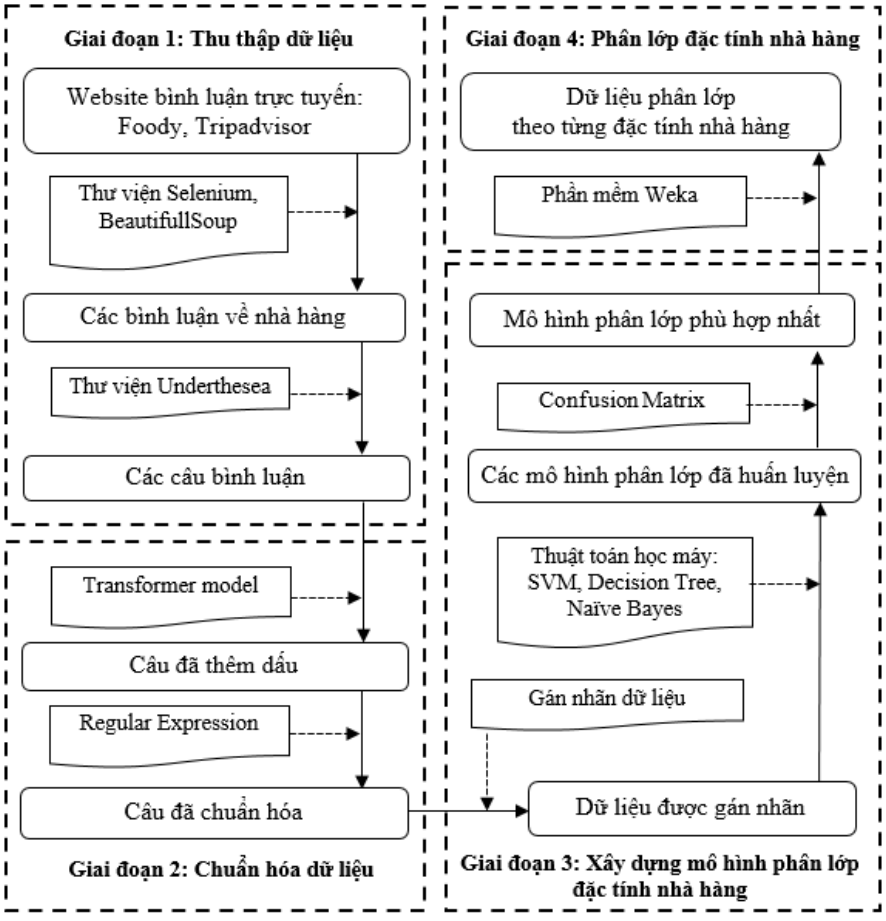
- Các từ chỉ mức độ trong tiếng Việt ảnh hưởng rất lớn đến việc đánh trọng số như: rất, cực, siêu, hơi, chưa, chả, khá, quá, chẳng.

Ngoài ra, khai phá quan điểm cũng gặp khó khăn khi văn bản chứa nhiều quan điểm và được đăng bởi nhiều người có quan điểm khác nhau.

Mô hình phân lớp chúng tôi đề xuất dựa vào phương pháp học máy nên sẽ giải quyết được thách thức về đặc tính ẩn của đối tượng. Ngoài ra, chúng tôi còn tiền xử lý dữ liệu bao gồm thêm dấu cho câu và chuẩn hóa láy âm tiết sẽ được trình bày cụ thể trong Mục 4.2. "Giai đoạn 2: Chuẩn hóa dữ liệu".

4 Mô hình phân lớp đặc tính nhà hàng sử dụng phương pháp học máy

Để vượt qua các thách thức trên, chúng tôi đề xuất một mô hình phân lớp đặc tính nhà hàng sử dụng phương pháp học máy được minh họa như trong Hình 1. Nguồn dữ liệu đưa vào phân tích là các bình luận thuộc lĩnh vực nhà hàng tại tỉnh Thừa Thiên Huế trên 2 website đánh giá trực tuyến Tripadvisor.com.vn và Foody.vn. Mô hình bao gồm 4 giai đoạn: (1) Thu thập dữ liệu (2) Chuẩn hóa dữ liệu (3) Xây dựng mô hình phân lớp đặc tính nhà hàng (4) Phân lớp đặc tính nhà hàng.



Hình 1. Mô hình phân lớp đặc tính nhà hàng sử dụng phương pháp học máy

4.1 Giai đoạn 1: Thu thập dữ liệu

Để thu thập dữ liệu là các bình luận của khách hàng từ các website đánh giá trực tuyến, chúng tôi sử dụng các thư viện của Python. Đầu tiên, thư viện Selenium và BeautifullSoup được sử dụng để thu thập các bình luận của khách hàng theo từng nhà hàng trên Tripadvisor.com.vn và Foody.vn [8]. Sau đó thư viện Underthesea [11] được dùng để thực hiện tách câu đối với những bình luận có nhiều hơn 2 câu. Thư viện Underthesea là rất hiệu quả để phân tích văn bản tiếng Việt trong các bài toán xử lý ngôn ngữ tự nhiên.

4.2 Giai đoạn 2: Chuẩn hóa 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, trước khi xây dựng mô hình phân lớp, chúng tôi tiền xử lý dữ liệu

bao gồm: thêm dấu cho câu và chuẩn hóa láy âm tiết. 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 [12]. Trong giai đoạn này, chúng tôi còn tiến hành chuẩn hóa dữ liệu tiếng Việt bằng cách sử dụng các kỹ thuật trong biểu thức chính quy (Regular Expression). Chúng tôi chuẩn hóa láy âm tiết đối với những từ thể hiện cảm xúc đặc biệt, như câu "*Món mực nướng ngonnn quá điiiiiiii!!!!!!!!*" sẽ được chuẩn hóa thành "*Món mực nướng ngon quá đi!" hoặc "Nhân viên phục vụ nhà hàng quá tuyệt vờiiiiiiii"* sẽ được chuẩn hóa thành *"Nhân viên phục vụ nhà hàng quá tuyệt vời*".

4.3 Giai đoạn 3: Xây dựng mô hình phân lớp đặc tính nhà hàng

Để xây dựng mô hình phân lớp đặc tính nhà hàng, chúng tôi thực hiện các bước: (a) Xác định các đặc tính của nhà hàng; (b) Gán nhãn cho bộ dữ liệu; và (c) Xây dựng mô hình theo thuật toán học máy.

a. Xác định các đặc tính của nhà hàng

Theo [13], [14], chất lượng thức ăn, chất lượng dịch vụ, giá cả và vị trí nhà hàng là các nhân tố có mức độ tác động mạnh nhất đến ý định lựa chọn nhà hàng. Dựa vào các nhân tố này, chúng tôi đề xuất các đặc tính thuộc lĩnh vực nhà hàng gồm: thức ăn (Food), dịch vụ (Service), giá cả (Price), môi trường xung quanh nhà hàng (Ambience), và nhà hàng (Restaurant). Các đặc tính của nhà hàng được mô tả chi tiết ở Bảng 1.

Bảng 1. Mô tả chi tiết các đặc tính của nhà hàng

Đặc tính	Mô tả chi tiết
Thức ăn	Thực đơn đa dạng, tốt cho sức khỏe, dinh dưỡng cao, thực phẩm được bảo quản ở nhiệt độ thích hợp, khẩu phần ăn chất lượng, trình bày món ăn hấp dẫn, thiết kế thực đơn phong phú, mùi vị đặc trưng của món ăn, độ tươi thực phẩm.
Dịch vụ	Nhà hàng quan tâm lợi ích của khách hàng, nhân viên luôn sẵn lòng giúp đỡ khách hàng, mọi thứ chu đáo, ngoại hình của nhân viên, nhân viên có kiến thức để trả lời câu hỏi của khách hàng, quản lý nhà hàng thân thiện.
Giá cả	Giá trị tổng thể của bữa ăn, thực đơn giá cả hợp lý, đáng đồng tiền.
Môi trường	Môi trường xung quanh, thiết kế bàn ăn, thẩm mỹ nhà hàng, trang trí, ánh sáng, bố cục, bãi đậu xe, vị trí nhà hàng.
Nhà hàng	Đề cập đến nhà hàng nói chung, khi không chỉ cụ thể một thực thể nào thì có nghĩa là đề cập chung đến nhà hàng.

b. Gán nhãn dữ liệu

Để có được dữ liệu xây dựng mô hình phân lớp, chúng tôi thu thập dữ liệu là tập các bình luận trên Tripadvisor.com.vn và Foody.vn. Qua quá trình thu thập, tiền xử lý dữ liệu, chúng tôi đã thu thập được 4.355 câu bình luận thuộc lĩnh vực nhà hàng. Sau đó, chúng tôi tiến hành gán nhãn dữ liệu bằng phương pháp thủ công cho 4.355 câu bình luận này. Để xây dựng mô hình phân lớp, chúng tôi sử dụng 3.955 câu bình luận, 400 câu bình luận còn lại sử dụng để kiểm chứng thực tế và đánh giá mô hình phân lớp. Các nhãn dữ liệu liên quan đến lĩnh vực nhà hàng bao gồm: thức ăn (Food), dịch vụ (Service), giá cả (Price), môi trường xung quanh nhà hàng (Ambience) và nhà hàng (Restaurant). Ví dụ về các câu bình luận đã gán nhãn được minh họa như trong Bảng 2.

Bảng 2. Các câu bình luận và nhãn dữ liệu

STT	Câu bình luận	Nhãn dữ liệu
1	Pizza ngon, đế bánh giòn, nóng hổi	Food
2	Nhân viên ở đây phục vụ nhanh và không phải chờ lâu	Service
3	Rẻ lại ngon, khá là hài lòng!!	Price
4	Quán rộng rãi, thoáng mát	Ambience
5	Đây là một trong những quán Pizza mình rất thích	Restaurant

c. Xây dựng mô hình theo thuật toán học máy

Trong học máy, máy tính không thể hiểu trực tiếp ngôn ngữ tự nhiên mà chỉ hiểu được ngôn ngữ khi chúng được biểu diễn dưới dạng không gian vector. Các chiều thuộc tính đầu vào sẽ được biểu diễn dưới dạng ma trận vector. Có nhiều phương pháp để biểu diễn văn bản sang dạng ma trận vector chẳng hạn như mô hình Bag of word N-grams hay mô hình TF-IDF (Term Frequency - Inverse Document Frequency). Tuy nhiên, mô hình Bag of word N-grams gặp một vài vấn đề đối với tập dữ liệu lớn, đó là các từ có tần suất xuất hiện nhiều ở đa số các đoạn văn bản, nhưng không có ý nghĩa phân lớp, ví dụ: như các từ "này", "đó", "rất"… Khi đó chỉ số TF-IDF sẽ được dùng để tính toán và phát hiện các từ có trọng số cao và thấp. Ngoài ra, thuật toán TF-IDF không biểu diễn giá trị của các thuộc tính bằng giá trị 0 và 1 mà sẽ biểu diễn với giá trị trọng số TF-IDF đã tính. Chính vì vậy, khi biểu diễn trên đồ thị giảm từ nhiều chiều sang 2 chiều, các giá trị của dữ liệu phân bổ phụ thuộc cả hai chiều, khi trục x tăng thay đổi thì cũng kéo theo giá trị trục y thay đổi. Do vậy dữ liệu phân bổ rời rạc và tách biệt hơn, việc này giúp quá trình phân lớp sẽ dễ dàng hơn [8]. Từ những lý do trên, chúng tôi sử dụng phương pháp TF-IDF để biểu diễn văn bản dưới dạng không gian vector trong thuật toán học máy.

Công thức tính TF-IDF (Term Frequency - Inverse Document Frequency)

Trọng số của từ (TF-IDF) thể hiện mức độ quan trọng của từ này trong một văn bản nằm trong một tập hợp các văn bản. Trọng số của từ dựa vào tần số xuất hiện của một từ trong một văn bản (TF) và tần số nghịch của một từ trong tập văn bản (IDF).

TF (Term Frequency) là tần số xuất hiện của một từ trong một văn bản. Giá trị TF của một từ t trong văn bản d được tính theo công thức (1).

$$tf(t,d) = \frac{n(t,d)}{n(d)} \quad (1)$$

trong đó, $n(t,d)$ là số lần xuất hiện từ t trong văn bản d; $n(d)$ là số lần xuất hiện của tất cả các từ trong văn bản d.

IDF (Inverse Document Frequency) là tần số nghịch của một từ trong tập văn bản. Trong tập văn bản, mỗi từ chỉ có một giá trị IDF duy nhất được tính theo công thức (2).

$$idf(t,D) = \log \frac{|D|}{|\{d \in D | t \in d\}|} \quad (2)$$

Trong đó, $|D|$ là tổng số văn bản trong tập D, $|\{d \in D | t \in d\}|$ là số văn bản d có xuất hiện từ t trong tập D.

TF-IDF được tính theo công thức (3)

$$\text{TF-IDF}(t,d,D) = \text{tf}(t, d) \times \text{idf}(t, D) \quad (3)$$

Trong đó, những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này và ít xuất hiện trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao.

Bước quan trọng nhất của quy trình phân lớp đặc tính nhà hàng là lựa chọn thuật toán phân lớp tốt nhất. Theo [15], bốn thuật toán đạt hiệu quả cao đối với trường hợp phân lớp văn bản theo đặc tính nhà hàng bao gồm: Naïve Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT) và K-Nearest Neighbor (KNN). Để huấn luyện mô hình phân lớp, chúng tôi sử dụng phần mềm Weka [16]. Kết quả huấn luyện mô hình được minh họa như trong Bảng 3.

Bảng 3. Kết quả huấn luyện mô hình

STT	Thuật toán	Precision	Recall	F1-Score	Thời gian (giây)
1	Naïve Bayes	0,699	0,675	0,682	4,20
2	Support Vector Machines	0,739	0,740	0,738	19,76
3	Decision Tree	0,671	0,672	0,671	170,37

STT	Thuật toán	Precision	Recall	F1-Score	Thời gian (giây)
4	K-Nearest Neighbor	0,705	0,485	0,504	0,01

Trong đó: Precision là độ chính xác, Recall là độ bao phủ, F1-Score là độ đo trung bình điều hòa.

Kết quả huấn luyện cho thấy mô hình huấn luyện sử dụng SVM cho kết quả tốt nhất với Precision = 0,739; Recall = 0,740 và F1-Score = 0,738. Dựa vào kết quả huấn luyện, chúng ta có thể kết luận mô hình sử dụng SVM là phù hợp với tập dữ liệu huấn luyện thuộc lĩnh vực nhà hàng.

4.4 Giai đoạn 4: Phân lớp đặc tính nhà hàng

Trong giai đoạn 3, chúng tôi đã xác định được mô hình phân lớp dựa trên SVM là phù hợp nhất đối với bài toán phân lớp đặc tính thuộc lĩnh vực nhà hàng. Chúng tôi dựa vào mô hình này để phân lớp đặc tính thuộc lĩnh vực nhà hàng đối với tập dữ liệu là các câu bình luận. Ngoài ra, để kiểm chứng hiệu quả phân lớp đặc tính nhà hàng của mô hình đã chọn, chúng tôi tiến hành thực nghiệm với tập dữ liệu gồm 400 câu bình luận đã được gán nhãn. Kết quả phân lớp đặc tính nhà hàng của mô hình đạt tỉ lệ phân lớp chính xác trung bình trên 78%.

5 Kết luận

Trong bài báo này, chúng tôi đã đề xuất một mô hình phân lớp đặc tính nhà hàng bằng phương pháp học máy. Để nâng cao độ chính xác cho kết quả phân lớp, chúng tôi đã áp dụng các kỹ thuật tiền xử lý dữ liệu gồm thêm dấu và chuẩn hóa láy âm tiết nhằm tăng ngữ nghĩa cho câu bình luận. Mô hình phân lớp đã xử lý được trường hợp đặc tính ẩn trong các câu bình luận nhằm giải quyết được thách thức đối với phương pháp từ vựng. Dựa trên kết quả thực nghiệm với bộ dữ liệu thuộc lĩnh vực nhà hàng, phương pháp học máy sử dụng thuật toán Support Vector Machines cho kết quả tốt nhất so với các thuật toán khác như Naïve Bayes, Decision Tree và K-Nearest Neighbor. 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 gia tăng đối tượng, phạm vi và dữ liệu phân tích. Ngoài ra, chúng tôi còn nghiên cứu thêm một số kỹ thuật tiền xử lý dữ liệu nhằm khắc phục được các thách thức trong việc chuẩn hóa văn bản tiếng Việt như: lỗi chính tả, sai ngữ pháp, dấu câu bị thiếu, chữ viết tắt và sử dụng tiếng lóng.

Tài liệu tham khảo

1. Le Thi Minh Nguyen (2019), *Text classification based on Support Vector Machine*, Dalat University Journal Of Science, Vol. 9, Issue 2, pp. 3–19.

2. Nguyễn Chí Hiếu (2022), *Khảo sát các mô hình phân loại văn bản tiếng Việt*, Tạp chí Khoa học và Công nghệ, Tập 03, Số 57, pp. 99-109.

3. B. C. Martínez-Seis, O. Pichardo-Lagunas, S. Miranda, I. J. Perez-Cazares, & J. A. Rodriguez-González (2022), *Deep Learning Approach for Aspect-Based Sentiment Analysis of Restaurants Reviews in Spanish*, *Computacion y Sistemas*, Vol. *26*, No. 2, pp. 899-908.

4. M. Kavin Prakash, S. Aravinth, D. Hari Nesha, M. Monica (2020), *Opinion Mining on Restaurant Rating Based on Aspects*, International Journal of Computational Science and Engineering, ISSN 2249-4251, Vol. 10, No. 1, pp. 25-34.

5. A. Suciati, I. Budi (2019), *Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia*, International Conference on Asian Language Processing (IALP), Shanghai, Nov 15-17, 2019.

6. Thái Kim Phụng, Nguyễn An Tế, Trần Thị Thu Hà (2020), *Hệ thống hỗ trợ đánh giá và khuyến nghị dịch vụ du lịch dựa trên khai thác ý kiến khách hàng trực tuyến*, Tạp chí Khoa học và Công nghệ, Tập 04, Số 46, pp. 175-189.

7. Nguyễn Thị Ngọc Tú, Nguyễn Đức Long, Nguyễn Khắc Giáo, Nguyễn Thị Thu Hà, Nguyễn Việt Anh (2017), *Một phương pháp phân tích quan điểm đánh giá của người dùng đối với chất lượng sản phẩm dựa trên các nhận xét cá nhân*, Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017.

8. Nguyễn Đặng Lập Bằng, Nguyễn Văn Hồ, Hồ Trung Thành (2020), *Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm*, Tạp chí Khoa học Đại học Mở Thành phố Hồ Chí Minh, Vol. 16, No. 1, pp. 64-78,.

9. K. Chitra, T. Kavitha, S. Hemalatha (2017), *A Survey on Opinion Mining: Techniques, Tools and Research Challenges in Sentiment Analysis*, Research Journal of Science and Engineering Systems, Vol. 1, pp. 43-51.

10. P. Haseena Rahmath (2014), *Opinion Mining and Sentiment Analysis challenges and Applications*, International Journal of Application or Innovation in Engineering & Management. Vol. 3, Issue 5, pp. 401-403.

11. Vu Anh (2019), *Underthesea - Vietnamese Natural Language Process Toolkit*, GNU General Public License Version 3.

12. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and L. Kaiser (2017), *Attention Is All You Need*, Proceedings of the 31st International Conference on Neural Information Processing System, Vol. 5, No. 11, pp. 6000-6010.

13. F. Nurifan, R. Sarno, K. R. Sungkono (2019), *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.

14. Nguyễn Thị Loan, Mai Anh Vũ, Hà Đình Hùng (2023), *Các nhân tố ảnh hưởng đến ý định lựa chọn nhà hàng của thực khách: nghiên cứu ứng dụng mô hình PLS-SEM*, Tạp chí Kinh tế và phát triển, số 310, pp. 84-93.

15. K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown (2019), *Text Classification Algorithms: A Survey*, Information (Switzerland), Vol. 10, No. 4, pp. 1-68.

16. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten (2009). *The WEKA data mining software: an update*, ACM SIGKDD explorations newsletter, Vol. 11, No. 1, pp. 10-18.