CLASSIFICATION OF SENTIMENTS USING SOME MACHINE LEARNING METHODS FOR VIETNAMESE TEXT

Le Van Hoa
School of Hospitality and Tourism – Hue University

ARTICLE INFO

Received: 29/8/2024
Revised: 13/11/2024
Published: 14/11/2024

KEYWORDS

Machine learning methods
Classification
Semantic
Online comments
Tourism

ABSTRACT

This paper uses several different machine learning methods to evaluate the sentiment classification capability for Vietnamese datasets. The datasets consist of online comments in the field of tourism. Additionally, the experiments compare and evaluate the sentiment classification results of the comments when applying semantic enhancement techniques for Vietnamese texts. The datasets used in the experiments were collected from Facebook fanpages in the field of tourism and online review websites such as Tripadvisor.com.vn and Foody.vn. The experiments use four machine learning algorithms: K-Nearest Neighbor, Support Vector Machines, Naïve Bayes, and Decision Tree. The results showed that the Support Vector Machines machine learning method provides the best sentiment classification performance compared to K-Nearest Neighbor, Naïve Bayes, and Decision Tree methods. This paper is valuable for sentiment classification applications in the field of tourism.

PHÂN LỚP CÁC QUAN ĐIỂM SỬ DỤNG MỘT SỐ PHƯƠNG PHÁP HỌC MÁY CHO VĂN BẢN TIẾNG VIỆT

Lê Văn Hòa
Trường Du lịch – ĐH Huế

THÔNG TIN BÀI BÁO

Ngày nhận bài: 29/8/2024
Ngày hoàn thiện: 13/11/2024
Ngày đăng: 14/11/2024

TỪ KHÓA

Phương pháp học máy
Phân lớp
Ngữ nghĩa
Bình luận trực tuyến
Du lịch

TÓM TẮT

Bài báo này sử dụng một số phương pháp học máy khác nhau để đánh giá khả năng phân lớp quan điểm cho dữ liệu tiếng Việt. Dữ liệu này là các câu bình luận trực tuyến về lĩnh vực du lịch. Ngoài ra, thực nghiệm còn so sánh và đánh giá kết quả phân lớp quan điểm các câu bình luận khi áp dụng các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt. Dữ liệu đưa vào thực nghiệm được thu thập từ các fanpage Facebook trong lĩnh vực du lịch và các website đánh giá trực tuyến như Tripadvisor.com.vn và Foody.vn. Thực nghiệm sử dụng 4 thuật toán học máy: K-Nearest Neighbor, Support Vector Machines, Naïve Bayes và Decision Tree. Kết quả cho thấy phương pháp học máy Support Vector Machines cho kết quả phân lớp quan điểm tốt nhất khi so sánh với các phương pháp K-Nearest Neighbor, Naïve Bayes, Decision Tree. Bài báo này có giá trị đối với các ứng dụng phân lớp quan điểm trong lĩnh vực du lịch.

DOI: https://doi.org/10.34238/tnu-jst.11035

Email: levanhoa84@hueuni.edu.vn

1. Giới thiệu

1.1. Bối cảnh nghiên cứu

Phân lớp văn bản là một lĩnh vực nghiên cứu quan trọng trong du lịch và khách sạn. Có ba kiểu phân lớp văn bản: đặc trưng, quan điểm và đánh giá. Việc phân lớp các bình luận và các đánh giá không chỉ giúp chúng ta biết được các đặc trưng văn bản mà còn biết được quan điểm và đánh giá của người tiêu dùng, điều này giúp các nhà nghiên cứu và quản lý du lịch hiểu rõ hành vi của người tiêu dùng và đánh giá chất lượng cũng như sự hài lòng của du khách [1], [2]. Ngoài ra, nghiên cứu trong [3] cho rằng mạng xã hội và các trang web đánh giá trực tuyến cho phép khách hàng đưa ra ý kiến của họ về sản phẩm hoặc dịch vụ thông qua các đánh giá, bình luận. Bằng cách sử dụng các đánh giá, bình luận trực tuyến của khách hàng về sản phẩm hoặc dịch vụ, khách hàng tiềm năng có thể chọn sản phẩm tốt nhất và các doanh nghiệp cũng có thể dễ dàng hiểu được hành vi mua hàng của khách hàng, cũng như sở thích và mức độ hài lòng của khách hàng về chất lượng sản phẩm hoặc dịch vụ. Đồng thời, khách hàng cũng cần thông tin tổng hợp ý kiến đánh giá của cộng đồng để có những quyết định chính xác. Chính vì thế, khai phá quan điểm đã trở thành một chủ đề hấp dẫn cho các nghiên cứu trong nhiều lĩnh vực khác nhau [4], [5].

Một nghiên cứu khác trong [6] cho rằng các bình luận trực tuyến do khách hàng đưa lên rất quan trọng đối với việc quản lý điểm đến du lịch vì nó có thể giúp họ thu thập các ý kiến phản hồi của khách hàng và đề xuất các giải pháp nhằm cải thiện hoạt động tổ chức du lịch. Có rất nhiều đánh giá, bình luận trên phương tiện truyền thông xã hội và các tổ chức này khó có thể phân tích chúng theo cách thủ công. Bằng cách áp dụng phân lớp quan điểm, các đánh giá có thể được phân thành nhiều lớp và giúp các tổ chức dễ dàng ra quyết định. Cũng theo nghiên cứu trong [7], để thực hiện phân lớp quan điểm, thông tin chủ quan cần được trích xuất từ các đánh giá, bình luận dưới dạng ngôn ngữ tự nhiên. Thông tin chủ quan này có thể là các quan điểm hoặc tình cảm. Để trích xuất quan điểm hoặc tình cảm từ các đánh giá, bình luận của khách hàng, có nhiều cách tiếp cận khác nhau như xử lý ngôn ngữ tự nhiên, phân tích văn bản, ngôn ngữ học tính toán và sinh trắc học. Chúng ta có thể sử dụng các phương pháp học máy khác nhau để phân lớp quan điểm vì chúng rất hiệu quả và dễ triển khai.

Trong phân lớp quan điểm, dữ liệu là một trong những yếu tố quan trọng đóng vai trò quyết định đến kết quả phân lớp. Đặc thù của nguồn dữ liệu lấy từ các website đánh giá trực tuyến và fanpage Facebook 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, sử dụng các kỹ thuật nhằm nâng cao ngữ nghĩa cho văn bản tiếng Việt để có bộ dữ liệu đạt chất lượng cao là một trong những nhiệm vụ quan trọng của bài toán phân lớp quan điểm. Ngoài ra, phương pháp phân lớp trong bài báo cho hiệu quả tốt trên miền dữ liệu du lịch.

1.2. Các nghiên cứu liên quan

Đã có một số nghiên cứu liên quan đến khả năng phân lớp quan điểm trong lĩnh vực du lịch. Cụ thể, nghiên cứu [8] tập trung vào việc phân tích, phân lớp các đánh giá và đã thử nghiệm đối với các đánh giá về lĩnh vực nhà hàng. Bài báo này tập trung vào khai phá quan điểm dựa trên khía cạnh cho các đánh giá bằng tiếng Tây Ban Nha. Nhóm tác giả cho rằng đánh giá trực tuyến về sản phẩm và dịch vụ đã trở nên quan trọng đối với khách hàng và doanh nghiệp. Tuy nhiên, nghiên cứu vẫn còn hạn chế như còn thiếu từ điển dữ liệu trong giai đoạn tiền xử lý để thay thế các từ quan trọng. Ngoài ra, nghiên cứu chưa tập trung xác định quan điểm dựa vào các biểu tượng cảm xúc. Một nghiên cứu khác trong [6], Haris và cộng sự đã đánh giá mức độ tích cực hoặc tiêu cực của các bình luận về điểm đến du lịch bằng cách sử dụng các bộ phân lớp SVM và Random Forest (RF), đồng thời tìm ra bộ phân lớp phù hợp cho tập dữ liệu du lịch. Tập dữ liệu được sử dụng cho nghiên cứu này là từ các bình luận về công viên Taman Negara. Hiệu suất của các bộ phân lớp này được so sánh nhằm mục đích xác định bộ phân lớp phù hợp nhất để sử dụng trong khai phá quan điểm đối với các đánh giá du lịch trên mạng xã hội. Trong nghiên cứu này,

SVM được chọn là bộ phân lớp phù hợp nhất để sử dụng với tập dữ liệu du lịch, đạt độ chính xác 67,97%, trong khi Random Forest đạt độ chính xác 63,55%. Tuy nhiên, những hạn chế gặp phải trong nghiên cứu này bao gồm không thể chạy nhiều mô hình do số lượng bản ghi nhỏ, thu thập bộ dữ liệu nhỏ và ở một điểm du lịch duy nhất nên dữ liệu chưa đa dạng.

Với các nghiên cứu trong nước về phân lớp quan điểm thuộc lĩnh vực du lịch, nghiên cứu trong [9] cho rằng phân lớp văn bản đóng vai trò quan trọng trong việc khai thác dữ liệu và phát hiện tri thức. Trong bài báo này, nhóm tác giả đã sử dụng các thuật toán như Naïve Bayes, SVM và K-NN để thực nghiệm phân lớp văn bản tiếng Việt trên 5 bộ dữ liệu thuộc 04 chủ đề khác nhau: Du lịch, giải trí, giáo dục và pháp luật. Kết quả thử nghiệm của nhóm tác giả cho thấy thuật toán SVM phân lớp văn bản tiếng Việt theo chủ đề là sự lựa chọn phù hợp cho các ứng dụng về phân lớp văn bản. Tuy nhiên, nhóm tác giả chưa sử dụng các kỹ thuật nâng cao ngữ nghĩa cho bộ dữ liệu tiếng Việt, trong khi chất lượng bộ dữ liệu sẽ ảnh hưởng rất lớn đến hiệu quả của mô hình phân lớp.

Một nghiên cứu khác trong [4] cho rằng 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. 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. Ngoài ra, nhóm tác giả 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, được thu thập tự động trên trang web Agoda. Tuy nhiên, nghiên cứu này vẫn còn nhiều hạn chế như: Nghiên cứu 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. Ngoài ra, một nghiên cứu trong [10] đã tiến hành thực nghiệm bằng phương pháp học máy để khai phá quan điểm từ các bình luận của khách hàng và trực quan hóa kết quả hỗ trợ ra quyết định. Nhóm tác giả đã thực nghiệm trên nhiều thuật toán để so sánh ưu và khuyết điểm của mô hình, từ đó chọn ra mô hình tốt nhất. Đóng góp của nhóm tác giả là đã đề xuất phương pháp khai phá quan điểm thông qua việc thu thập tập dữ liệu là ý kiến bình luận của khách hàng trên website Foody.vn. 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.

Trong bài báo này, chúng tôi đề xuất một mô hình phân lớp quan điểm các câu bình luận tiếng Việt với nhiều ưu điểm nổi bật. Thứ nhất, bộ dữ liệu được thu thập từ nhiều nguồn khác nhau, đảm bảo tính phong phú và đa dạng. Thứ hai, chất lượng của bộ dữ liệu được cải thiện thông qua việc áp dụng các kỹ thuật nâng cao ngữ nghĩa dành riêng cho văn bản tiếng Việt. Cuối cùng, mô hình phân lớp đã được thực nghiệm trên nhiều phương pháp học máy khác nhau, đảm bảo tính tổng quát và hiệu quả của kết quả nghiên cứu.

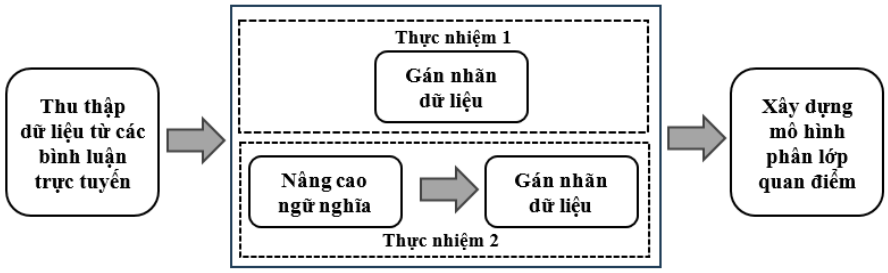
2. Phương pháp nghiên cứu

2.1. Đối tượng nghiên cứu

Nghiên cứu này tập trung chủ yếu vào việc phân lớp quan điểm các câu bình luận tiếng Việt thuộc lĩnh vực du lịch sử dụng một số phương pháp học máy. Các câu bình luận tiếng Việt sử dụng để phân lớp quan điểm là các bình luận thuộc lĩnh vực du lịch trên các fanpage Facebook du lịch và 2 website đánh giá trực tuyến Tripadvisor.com.vn và Foody.vn.

2.2. Mô hình nghiên cứu đề xuất

Để phân lớp quan điểm các câu bình luận, chúng tôi dựa vào các nghiên cứu [6], [7], [11] nhằm đề xuất một mô hình phân lớp quan điểm các câu bình luận tiếng Việt sử dụng một số phương pháp học máy áp dụng cho miền dữ liệu thuộc lĩnh vực du lịch được minh họa trong Hình 1.



Hình 1. *Mô hình phân lớp quan điểm các câu bình luận tiếng Việt sử dụng một số phương pháp học máy áp dụng cho miền dữ liệu thuộc lĩnh vực du lịch*

Mô hình bao gồm 4 giai đoạn: (1) Thu thập dữ liệu; (2) Nâng cao ngữ nghĩa cho văn bản tiếng Việt; (3) Gán nhãn dữ liệu; (4) Xây dựng mô hình phân lớp quan điểm. Trong mô hình này, để so sánh và đánh giá kết quả phân lớp quan điểm các câu bình luận trước và sau khi áp dụng các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt, chúng tôi tiến hành 2 thực nghiệm: (1) Thực nghiệm 1 đối với dữ liệu chưa qua giai đoạn nâng cao ngữ nghĩa cho văn bản tiếng Việt; (2) Thực nghiệm 2 đối với dữ liệu đã qua giai đoạn nâng cao ngữ nghĩa cho văn bản tiếng Việt.

2.2.1. Phương pháp thu thập dữ liệu

Nguồn dữ liệu đưa vào thực nghiệm được thu thập từ các fanpage Facebook trong lĩnh vực du lịch và các website đánh giá trực tuyến như Tripadvisor.com.vn và Foody.vn. Để thu thập dữ liệu là các bình luận trực tuyến theo từng nhà hàng từ các website đánh giá trực tuyến như Tripadvisor.com.vn và Foody.vn, chúng tôi sử dụng bộ thư viện 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 [10]. Ngoài ra, để thu thập dữ liệu từ các fanpage Facebook trong lĩnh vực du lịch, chúng tôi sử dụng bộ công cụ RestFB là một Java API giúp tương tác với Facebook Graph API. API này cho phép thu thập thông tin từ các thực thể, như các bài viết (post), các bình luận (comment) [12]. Sau đó chúng tôi sử dụng thư viện Underthesea [13] 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.

2.2.2. Các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt

Với đặc thù dữ liệu đầu vào của mô hình này là các câu bình luận trực tuyến về lĩnh vực du lịch nên người viết bình luận không theo một chuẩn nhất định. Tùy theo trình độ học vấn, giới tính, độ tuổi mà người đăng bình luận có cách hành văn khác nhau. Để nâng cao ngữ nghĩa cho các câu bình luận, chúng tôi tiến hành các kỹ thuật: thêm dấu cho câu đối với các câu tiếng Việt không dấu; chuẩn hóa láy âm tiết; chuẩn hóa chữ viết tắt; xử lý biểu tượng cảm xúc.

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]. Kiến trúc mô hình dịch máy sử dụng Transformer gồm hai phần là Encoder và Decoder. Encoder biểu diễn chuỗi đầu vào là văn bản không dấu thành một chuỗi vector ẩn với ngữ nghĩa và thông tin ngữ cảnh đầy đủ. Decoder giải mã chuỗi vector ẩn từ Encoder thành chuỗi đầu ra là văn bản có dấu. Ngoài ra, chúng tôi còn tiến hành nâng cao ngữ nghĩa cho văn bản tiếng Việt sử dụng các kỹ thuật trong biểu thức chính quy (Regular Expression). Trường hợp thứ nhất: chuẩn hóa láy âm tiết, như câu "Điểm du lịch này quá tuyệt vờiiiiiiii !!!!!!!!" sẽ được chuẩn hóa thành "Điểm du lịch này quá tuyệt vời !" hoặc "Nhiều món ăn ngon quá điiiiiiiii" sẽ được chuẩn hóa thành "Nhiều món ăn ngon quá đi". Trường hợp thứ hai: chuẩn hóa chữ viết tắt, hệ thống thực hiện việc thay thế các từ như: "khong", "ko" thành từ "không" hoặc "đc", "dc" thành từ "được". Một kỹ thuật khác, chúng tôi dựa vào công cụ Demojize [15] để xử lý biểu tượng cảm

xúc bằng cách chuyển các biểu tượng cảm xúc này thành văn bản. Sau khi áp dụng các kỹ thuật này, chúng tôi thu thập được các câu bình luận đã chuẩn hóa và xử lý biểu tượng cảm xúc.

2.2.3. Phương pháp gán nhãn dữ liệu

Trong quá trình nghiên cứu, chúng tôi đã tiến hành thu thập và tiền xử lý một tập dữ liệu bao gồm 14.900 câu bình luận tiếng Việt trong lĩnh vực du lịch. Để đảm bảo không trùng lặp giữa tập dữ liệu huấn luyện (train) và tập dữ liệu kiểm tra (test), chúng tôi tiến hành loại bỏ các câu bình luận trùng lặp. Sau khi loại bỏ 593 câu trùng lặp từ tập dữ liệu 14.900 câu bình luận, tập dữ liệu cuối cùng gồm 14.307 câu bình luận, các câu bình luận này được gán nhãn thủ công để phục vụ cho mục đích xây dựng mô hình phân lớp. Cụ thể, dữ liệu gán nhãn được phân thành ba lớp: lớp bình luận tích cực được gán nhãn là 1, lớp bình luận trung lập được gán nhãn là 0, và lớp bình luận tiêu cực được gán nhãn là -1. Trong tổng số 14.307 câu bình luận, có 9.751 câu được gán nhãn tích cực, 1.063 câu được gán nhãn trung lập và 3.493 câu được gán nhãn tiêu cực. Ví dụ về các câu bình luận đã được gán nhãn như trong Bảng 1.

Bảng 1. *Các câu bình luận và nhãn dữ liệu*

STT	Câu bình luận	Nhãn dữ liệu
1	Kiến trúc độc đáo, cổ kính	1
2	Chùa cổ đẹp lắm mọi người ạ	1
3	Nhân viên ở đây phục vụ nhanh và không phải chờ lâu	1
4	Dịch vụ cũng bình thường	0
5	Bánh gạo cay ăn cũng tạm	0
6	Bình thường, nói chung nhác nấu thì có thể ra đây ăn tối	0
7	Vệ sinh không đảm bảo	-1
8	Không gian nhà hàng chật hẹp	-1
9	Nhiều khu vực bị xuống cấp chưa được trùng tu	-1

2.2.4. Phương pháp xây dựng mô hình phân lớp quan điểm

Để xây dựng mô hình phân lớp quan điểm, chúng tôi tiến hành phân chia tỉ lệ giữa bộ dữ liệu huấn luyện và bộ dữ liệu kiểm tra từ bộ dữ liệu thu thập được. Việc phân chia bộ dữ liệu rất quan trọng vì được kỳ vọng sẽ tạo ra bộ dữ liệu kiểm tra tốt nhất để đánh giá hiệu suất của mô hình phân lớp dựa trên trên bộ dữ liệu huấn luyện. Trong mô hình này, chúng tôi dựa vào nghiên cứu [16] để chia bộ dữ liệu thu thập được theo tỉ lệ: 80% cho bộ dữ liệu huấn luyện và 20% cho bộ dữ liệu kiểm tra. Như vậy, trong 14.307 câu bình luận đã được gán nhãn, chúng tôi sử dụng 11.446 câu bình luận làm bộ dữ liệu huấn luyện, 2.861 câu bình luận còn lại làm bộ dữ liệu kiểm tra.

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. Trong mô hình này, chúng tôi sử dụng phương pháp TF-IDF (Term Frequency - Inverse Document Frequency) để biểu diễn văn bản dưới dạng không gian vector bởi vì phương pháp này có những ưu điểm như: có khả năng loại bỏ các từ không có ý nghĩa phân lớp; biểu diễn trên đồ thị giảm từ nhiều chiều sang 2 chiều; dữ liệu phân bổ rời rạc và tách biệt nên quá trình phân lớp sẽ dễ dàng hơn [10].

Bước quan trọng nhất của việc xây dựng mô hình phân lớp quan điểm là lựa chọn phương pháp phân lớp tốt nhất. Theo [17], bốn phương pháp đạt hiệu quả cao đối với trường hợp phân lớp văn bản bao gồm: K-Nearest Neighbor, Support Vector Machines, Naïve Bayes và Decision Tree. Một số đặc điểm cơ bản của bốn phương pháp phân lớp:

- Phương pháp K-Nearest Neighbors (KNN): Ý tưởng cơ bản để phân lớp là dựa vào khoảng cách gần nhất giữa đối tượng cần phân lớp và tất cả các đối tượng trong tập dữ liệu huấn luyện. Một đối tượng được phân lớp dựa vào K láng giềng của nó. Hiệu quả phân lớp của KNN phụ thuộc nhiều vào các tham số, trong đó K được coi là tham số quan trọng nhất. Các giá trị K nhỏ (ví dụ: 1, 2 hoặc 3) thường cho hiệu suất tốt trong những trường hợp dữ liệu ít nhiễu, trong khi các giá trị K lớn hơn (ví dụ: 5, 6, 7 hoặc 10) có thể giảm thiểu nguy cơ overfitting. Hơn nữa,

trong KNN mặc dù nhiều thước đo khoảng cách có thể được sử dụng, khoảng cách Euclid thường hiệu quả hơn so với các thước đo khác như khoảng cách Manhattan và Chebyshev.

- Phương pháp Support Vector Machines (SVM): Ý tưởng của SVM là tìm một siêu phẳng để phân tách các điểm dữ liệu. Siêu phẳng này sẽ chia không gian thành các miền khác nhau và mỗi miền sẽ chứa một loại dữ liệu. Hiệu quả của việc phân lớp bằng SVM phụ thuộc nhiều vào các tham số quan trọng. Một trong số đó là tham số C (Complexity), có vai trò kiểm soát mức độ chấp nhận sai số trên tập huấn luyện, từ đó giúp cân bằng giữa việc tối đa hóa biên độ phân tách và việc giảm thiểu sai số trong quá trình huấn luyện. Bên cạnh đó, việc xác định hàm nhân cũng đóng vai trò quan trọng trong hiệu suất của SVM. Cụ thể, hàm nhân tuyến tính (Linear Kernel) được sử dụng khi dữ liệu có thể phân tách tuyến tính, thường cho hiệu suất tốt hơn trong các trường hợp dữ liệu lớn với cấu trúc đơn giản. Hàm nhân đa thức (Polynomial Kernel) lại phù hợp hơn cho các dữ liệu có tính phi tuyến đơn giản. Trong khi đó, hàm nhân RBF (Radial Basis Function) thường hoạt động hiệu quả với dữ liệu có tính phi tuyến cao. Thêm vào đó, tham số Gamma trong RBF kiểm soát độ ảnh hưởng của một điểm dữ liệu đơn lẻ đến các điểm lân cận, góp phần điều chỉnh độ nhạy của mô hình đối với sự thay đổi trong dữ liệu.

- Phương pháp Naïve Bayes (NB): Ý tưởng cơ bản của phương pháp NB là sử dụng xác suất có điều kiện giữa các đặc trưng và nhãn để dự đoán xác suất nhãn của văn bản cần phân lớp. Điểm quan trọng của phương pháp này chính là giả định rằng sự xuất hiện của tất cả các đặc trưng trong văn bản đều độc lập với nhau. Hiệu quả của phương pháp phân lớp NB phụ thuộc vào một số tham số quan trọng. Một trong số đó là, tham số Kernel Estimator sẽ được lựa chọn để cải thiện độ chính xác của mô hình khi dữ liệu không tuân theo phân phối chuẩn. Tham số thứ hai là useSupervisedDiscretization, cho phép mô hình tự động điều chỉnh và cải thiện khả năng xử lý các thuộc tính rời rạc một cách hiệu quả.

- Phương pháp Decision Tree (DT): Cây quyết định là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Khi cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết. Hiệu quả của phương pháp phân lớp bằng Decision Tree phụ thuộc vào nhiều tham số quan trọng. Một trong số đó là yếu tố Confidence Factor (CF), xác định mức độ cắt tỉa của cây. Cụ thể, giá trị CF cao sẽ dẫn đến việc cây ít bị cắt tỉa, trong khi giá trị CF thấp sẽ khiến cây bị cắt tỉa nhiều hơn. Tham số thứ hai là MinNumObj, xác định số lượng mẫu tối thiểu trong mỗi nút lá. Giá trị nhỏ cho tham số này có thể tạo ra các nhánh nhỏ, trong khi giá trị lớn hơn có thể làm đơn giản hóa cây quá mức; giá trị mặc định thường được thiết lập là 2.

Chúng tôi sử dụng phần mềm Weka [18] để xây dựng và huấn luyện các mô hình phân lớp các quan điểm dựa trên bốn phương pháp học máy. Trong quá trình xây dựng các mô hình này, các tham số của từng phương pháp được tinh chỉnh thông qua việc điều chỉnh các thiết lập trong Weka nhằm tối ưu hóa hiệu quả phân lớp. Cụ thể, đối với phương pháp KNN, giá trị của tham số K đã được thay đổi, và kết quả cho thấy khi K = 3 và sử dụng hàm khoảng cách Euclid, độ chính xác đạt được là cao nhất. Đối với phương pháp SVM, hàm nhân RBF với giá trị Gamma = 0,01 và tham số C = 10 đã mang lại kết quả tốt nhất. Trong các thử nghiệm với phương pháp Naïve Bayes, việc thiết lập Kernel Estimator = True cho thấy hiệu quả vượt trội. Cuối cùng, đối với phương pháp Decision Tree, các tham số CF = 0,25 và MinNumObj = 2 được xác định là tối ưu sau quá trình thử nghiệm.

3. Kết quả nghiên cứu

Các mô hình phân lớp quan điểm được huấn luyện và đánh giá thông qua bộ dữ liệu là 11.446 câu bình luận tiếng Việt thuộc lĩnh vực du lịch. Các mô hình này được huấn luyện trên máy tính chạy hệ điều hành Windows 10, cấu hình máy Processor Intel Core i5 tốc độ 26,7 GHz, RAM 8 GB, ổ cứng 500 GB. Kết quả huấn luyện mô hình phân lớp quan điểm như trong Bảng 2.

Bảng 2. *Kết quả huấn luyện mô hình phân lớp quan điểm*

STT	Phương pháp	Thực nghiệm 1: Bộ dữ liệu chưa qua giai đoạn nâng cao ngữ nghĩa				Thực nghiệm 2: Bộ dữ liệu đã qua giai đoạn nâng cao ngữ nghĩa			
		Precision	Recall	F1-Score	Thời gian (s)	Precision	Recall	F1-Score	Thời gian (s)
1	KNN (K = 3; hàm Euclid)	0,659	0,690	0,652	0,02	0,687	0,707	0,637	0,02
2	SVM (C = 10; RBF; Gamma = 0,01)	0,781	0,781	0,748	330,05	0,814	0,825	0,814	247,31
3	NB(Kernel Estimator = True)	0,727	0,687	0,699	7,53	0,752	0,762	0,755	7,6
4	DT (CF = 0,25; MinNumObj = 2)	0,701	0,726	0,714	732,5	0,726	0,745	0,733	579,58

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 mô hình phân lớp quan điểm ở Bảng 2 cho thấy mô hình huấn luyện sử dụng phương pháp SVM cho kết quả tốt nhất ở thực nghiệm 2 (Bộ dữ liệu đã qua giai đoạn nâng cao ngữ nghĩa cho văn bản tiếng Việt) với Precision = 0,814; Recall = 0,825 và F1-Score = 0,814. Phương pháp SVM đạt kết quả phân lớp vượt trội so với các phương pháp khác như KNN, NB và DT trong bài toán phân lớp quan điểm văn bản tiếng Việt, bởi vì một số lý do chính. Thứ nhất, SVM có khả năng xử lý dữ liệu không tuyến tính thông qua việc sử dụng hàm nhân RBF, giúp biến đổi không gian đặc trưng ban đầu thành không gian có chiều cao hơn. Điều này cho phép SVM xử lý hiệu quả các tập dữ liệu phức tạp và không tuyến tính, đặc biệt là trong các câu bình luận tiếng Việt thuộc lĩnh vực du lịch. Thứ hai, việc tinh chỉnh các tham số như Gamma và C trong SVM có vai trò quan trọng trong tối ưu hóa quá trình học và nâng cao độ chính xác của mô hình. Như vậy, trong phân lớp quan điểm việc áp dụng các kỹ thuật nâng cao ngữ nghĩa đã cải thiện đáng kể độ chính xác của mô hình. Dựa trên các phân tích và thực nghiệm, chúng tôi kết luận rằng mô hình sử dụng phương pháp SVM phù hợp nhất với tập dữ liệu huấn luyện trong lĩnh vực du lịch, đặc biệt sau khi áp dụng các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt.

4. Kết luận

Trong bài báo này, chúng tôi đã xác định được mô hình phân lớp quan điểm phù hợp nhất bằng cách so sánh, đánh giá kết quả phân lớp quan điểm dựa vào một số phương pháp học máy khi áp dụng các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt. Từ kết quả huấn luyện mô hình phân lớp quan điểm, chúng tôi có thể kết luận mô hình sử dụng phương pháp SVM là phù hợp nhất với bộ dữ liệu huấn luyện thuộc lĩnh vực du lịch và sau khi áp dụng các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt. Mô hình phân lớp quan điểm 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. Trong bài báo này, chúng tôi đã gia tăng tỉ lệ phân lớp quan điểm văn bản tiếng Việt từ độ chính xác Precision = 0,781; Recall = 0,781 và F1-Score = 0,748 lên độ chính xác Precision = 0,814; Recall = 0,825 và F1-Score = 0,814 đối với 11.446 câu bình luận sau khi áp dụng tất cả các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt.

Mô hình phân lớp quan điểm chúng tôi đề xuất có thể được ứng dụng để xây dựng các hệ thống liên quan đến tư vấn hoặc khai phá quan điểm trong lĩnh vực du lịch. Trong thời gian tới, chúng tôi tiếp tục tìm hiểu thêm các kỹ thuật nâng cao ngữ nghĩa cho văn bản tiếng Việt nhằm tăng khả năng phân lớp quan điểm cho mô hình.

Lời cám ơn

Nghiên cứu này được tài trợ bởi đề tài Khoa học và Công nghệ cấp Đại học Huế, Mã số đề tài: DHH2023-10-30, tên đề tài: "Xây dựng hệ thống tư vấn dựa trên bình luận trực tuyến cho một số nhà hàng tại Thừa Thiên Huế".

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] J. Liu, S. Hu, F. Mehraliyev, and H. Liu, "Text classification in tourism and hospitality – a deep learning perspective," *International Journal of Contemporary Hospitality Management*, vol. 35, no. 12, pp. 4177-4190, 2023.

[2] L. Bharadwaj, "Sentiment Analysis in Online Product Reviews: Mining Customer Opinions for Sentiment Classification," *International Journal for Multidisciplinary Research*, vol. 5, no. 5, pp. 1-34, 2023.

[3] A. Kohakade, G. Jogdand, P. Kohokade, K. Kadam, and P. A. Kadam, "A Machine Learning Approach for Opinion Mining Online Customer Reviews," *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 3, no. 12, pp. 108-111, 2023.

[4] K. P. Thai, A. T. Nguyen, and T. T. H. Tran, "A support system for tourism services assessment and recommendation based on pinion mining online customer reviews," (in Vietnamese), *Journal of Science and Technology - Industrial University of HCMC*, vol. 46, no. 4, pp. 175-189, 2020.

[5] A. Suciati and I. Budi, "Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia," in *International Conference on Asian Language Processing*, Shanghai, 2019, pp. 59-64.

[6] N. A. K. M. Haris, S. Mutalib, A. M. A. Malik, S. Abdul-Rahman, and S. N. K. Kamarudin, "Sentiment classification from reviews for tourism analytics," *International Journal of Advances in Intelligent Informatics*, vol. 9, no. 1, pp. 108-120, 2023.

[7] S. K. N. Prasanthi, M. C. S. Rao, and S. Chekuri, "Sentiment mining of customer reviews from e-commerce websites," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 6, pp. 2401-2407, 2024.

[8] B.-C. Martınez-Seis, O. Pichardo-Lagunas, S. Miranda, I.-J. Perez-Cazares, and J.-A. Rodriguez-Gonzalez, "Deep Learning Approach for Aspect-Based Sentiment Analysis of Restaurants Reviews in Spanish," *Computacion y Sistemas*, vol. 26, no. 2, pp. 899-908, 2022.

[9] T. L. Manh, V. V. Vu, V. L. Nguyen, T. H. M. Lam, T. T. T. Nguyen, and T. M. T. Duong, "Automatically Vietnamese text classification by topic," (in Vietnamese), *Journal of Science: Technology and Food - Ho Chi Minh City University of Food Industry*, vol. 18, no. 1, pp. 129-139, 2019.

[10] D. L. B. Nguyen, V. H. Nguyen, and T. T. Ho, "A text-based model for opinion mining and sentiment analysis from online customer reviews in food industry," (in Vietnamese), *Ho Chi Minh City Open University Journal of Science*, vol. 16, no. 1, pp. 64-78, 2020.

[11] B. Kejiya and S. K. Alisha, "Machine Learning Approach for Opinion Mining Online Customer Review," *International Journal of Engineering Science and Advanced Technology (IJESAT)*, vol. 24, no. 5, pp. 80-86, 2024.

[12] K. Jasmeet and S. Neha, "Facebook Integration with RESTFB API," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 3, no. 11, pp. 3891-3894, 2014.

[13] A. Vu, *Underthesea - Vietnamese Natural Language Process Toolkit.* [Performance]. GNU General Public License Version 3, 2019.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and Ł. Kaiser, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing System*, California, 2017, pp. 6000-6010.

[15] T. Kim and K. Wurster, *emoji v.0.3.4.* [Performance]. BSD License, 2015.

[16] I. O. Muraina, "Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts," in *7th International Mardin Artuklu Scientific Researches Conference*, Mardin, 2022, pp. 496-504.

[17] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," *Journal of Information*, vol. 10, no. 4, pp. 1-68, 2019.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18, 2009.