Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế-Luật và Quản lý 2023, 7(1):4089-4103

Open Access Full Text Article

Bài nghiên cứu

Phân tích cảm xúc và hành vi người dùng trực tuyến trong lĩnh vực du lịch tại Việt Nam dựa vào đánh giá và nội dung bình luận

Hồ Trung Thành[1,2], Nguyễn Văn Hồ[1,2], Lê Hoành Sử[1,2], Lê Thị Kim Hiền[1,2], Nguyễn Quang Phúc[1,2], Lê Bá Thiền[1,2,*]

Use your smartphone to scan this QR code and download this article

TÓM TẮT

Là một điểm đến hấp dẫn, ngành du lịch Việt Nam thu hút lượng lớn khách du lịch và trở thành ngành mũi nhọn đóng vai trò quan trọng trong tổng sản phẩm nội địa (GDP) của đất nước. Lượng khách du lịch lớn kéo theo sự phát triển của lĩnh vực khách sạn, đặc biệt là dịch vụ kinh doanh khách sạn trực tuyến. Ngày nay, sự bùng nổ của các nền tảng xã hội và các trang thương mại điện tử đã giúp doanh nghiệp có lượng lớn dữ liệu thu từ các phản hồi của khách hàng. Khai thác được nguồn tài nguyên vô tận này sẽ giúp doanh nghiệp có lợi thế cạnh tranh so với các đối thủ. Trong phạm vi nghiên cứu này, phân tích quan điểm trong cảm xúc của khách hàng dựa trên những đánh giá, bình luận được tập trung xem xét. Phân tích cảm xúc rất hữu ích giúp cho nhà quản trị doanh nghiệp nắm bắt và hiểu rõ nhu cầu, mong muốn của khách hàng. Từ đó, đưa ra các chiến lược phù hợp để phát triển và cải thiện chất lượng sản phẩm, dịch vụ. Dữ liệu của bài nghiên cứu được thu thập trên nền tảng du lịch trực tuyến agoda.com với 272.835 bình luận. Nghiên cứu sử dụng mô hình học máy và các thuật toán Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF) và Naïve Bayes (NB) để đánh giá và phân tích quan điểm khách hàng trong lĩnh vực kinh doanh khách sạn trực tuyến tại Việt Nam. Kết quả nghiên cứu có độ chính xác cao lên đến 90% ở cả hai loại ngôn ngữ tiếng Anh và tiếng Việt.

Từ khoá: bình luận khách hàng, phân tích cảm xúc, phân tích hành vi, máy học

GIỚI THIỆU

Những năm gần đây, với sự phát triển mạnh mẽ của các công nghệ mới đã thúc đẩy sự tăng trưởng nhanh chóng của các nền tảng truyền thông trên Internet và làm xuất hiện hình thức truyền miệng trực tuyến hay còn gọi là truyền miệng điện tử (eWOM). eWOM có thể được thể hiện dưới nhiều hình thức khác nhau như ý kiến, xếp hạng trực tuyến, phản hồi trực tuyến, đánh giá, bình luận, chia sẻ kinh nghiệm trên các kênh truyền thông trực tuyến [1]. Ngày càng có nhiều người tiêu dùng đưa ra quyết định mua hàng dựa trên các đánh giá của những khách hàng trước [2]. Theo khảo sát từ đánh giá của người tiêu dùng của BrightLocal, 91% người tiêu dùng trong độ tuổi 18-34 tin tưởng các đánh giá trực tuyến và xem xét đó là thông tin tham khảo để tiến hành mua hàng hay sử dụng dịch vụ. Đối với lĩnh vực du lịch và lữ hành, khách du lịch trong thời đại ngày nay thường có xu hướng xem xét các ý kiến, hình ảnh, đánh giá khách sạn từ những người đã sử dụng dịch vụ để lên kế hoạch cho bản thân. Những đánh giá trực tuyến như vậy đã ảnh hưởng đến hơn 10 tỷ đô la hàng năm trong việc đặt dịch vụ du lịch trực tuyến [3]. Các nền tảng trực tuyến đang tạo ra kiến thức tập thể và trở thành nguồn thu thập thông tin chính của khách du lịch khi đưa ra quyết định du lịch và mua các sản phẩm và dịch vụ liên quan [3].

Theo số liệu từ nghiên cứu [4,5], trong 80% khách du lịch tìm kiếm khách sạn trên trang web, sẽ có hơn 50% trong số họ đặt phòng qua trang web hoặc ứng dụng. Do đó, nếu khách sạn đưa ra một số tiêu chí theo sở thích của khách hàng thì sẽ giúp khách hàng dễ dàng tìm kiếm dịch vụ thuận tiện hơn [6]. Từ đó, nắm bắt được bức tranh chính xác và đầy đủ về khách hàng là một nhiệm vụ đầy thách thức đối với doanh nghiệp trong ngành du lịch nói riêng và tất cả các ngành nghề nói chung [7]. Ngày này, với sự phát triển của các mạng xã hội và cổng thông tin du lịch, lĩnh vực khách sạn đã có những bước phát triển mạnh mẽ, số lượng các đánh giá của khách hàng trực tuyến ngày càng tăng và đóng vai trò quan trọng [8]. Để dễ dàng tiếp cận và cạnh tranh với thị trường toàn cầu, ngành khách sạn Việt Nam cần hướng tới các tiêu chuẩn quốc tế. Các nhà quản lý khách sạn Việt Nam cũng cần có phương pháp đo lường mức độ hài lòng của khách, từ đó nâng cao chất lượng dịch vụ để thu hút thêm nhiều khách từ nhiều quốc gia trên thế giới đến với Việt Nam [6].

Trong giai đoạn chuyển tiếp và thích nghi với nền kinh tế số, cần thiết phải có một hướng tiếp cận mới về phân tích trải nghiệm, quan điểm người dùng để đem

[1]*Trường Đại học Kinh tế - Luật*

[2]*Đại học Quốc gia Thành phố Hồ Chí Minh*

Liên hệ

Lê Bá Thiền, Trường Đại học Kinh tế - Luật Đại học Quốc gia Thành phố Hồ Chí Minh

Email: thienlb.ktl@uel.edu.vn

Lịch sử

- Ngày nhận: 02-02-2023
- Ngày chấp nhận: 24-3-2023
- Ngày đăng: 31-3-2023

DOI :

https://doi.org/10.32508/stdjelm.v7i1.1187

Check for updates

VNU-HCM Press

Trích dẫn bài báo này: Thành H T, Hồ N V, Sử L H, Hiền L T K, Phúc N Q, Thiền L B. **Phân tích cảm xúc và hành vi người dùng trực tuyến trong lĩnh vực du lịch tại Việt Nam dựa vào đánh giá và nội dung bình luận**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 7(1):4089-4103.

lại hiệu quả trong việc dự đoán và tận dụng những công nghệ mang tính đột phá. Những tiến bộ của công nghệ thông tin đã làm thay đổi cách thức truyền thông giúp cho khách hàng dễ dàng truy cập thông tin và trao đổi ý kiến về sản phẩm và dịch vụ trên một quy mô lớn trong thời gian thực[9]. Bên cạnh đó, phát hiện ra quan điểm người dùng đúng sẽ giúp các nhà tiếp thị đưa ra nội dung quảng cáo tốt hơn[10]. Nhìn chung, phân tích quan điểm là việc khai thác các đánh giá trực tuyến của khách hàng, để cập đến việc áp dụng các kỹ thuật học máy để đánh giá, phân loại thái độ và quan điểm về một chủ đề cụ thể. Đây là một xu hướng nghiên cứu mới giúp các nhà quản lý hiểu rõ trải nghiệm cũng như mức độ hài lòng khách hàng khi sử dụng sản phẩm, dịch vụ mà doanh nghiệp cung cấp[6]. Ngoài ra, sự hiện diện của các công nghệ được xây dựng dựa trên trí tuệ nhân tạo (Artificial Intelligence - AI) và các phương pháp học máy (Machine Learning - ML) cùng các công cụ phân tích dữ liệu giúp cho việc phân tích và trích xuất dữ liệu dưới dạng văn bản (các bình luận, phản hồi của khách hàng) được dễ dàng và thiết thực hơn.

Bài nghiên cứu này trình bày mô hình phân tích quan điểm dựa trên các đánh giá của khách hàng đối với doanh nghiệp kinh doanh trong lĩnh vực khách sạn trên nền tảng trực tuyến. Kết quả bài nghiên cứu giúp các công ty hiểu rõ những mong muốn của khách hàng đối với sản phẩm và dịch vụ mà họ cung cấp. Những người chủ doanh nghiệp không chỉ muốn nhận được đánh giá tốt mà còn muốn tìm hiểu tốt nhất về khách hàng của mình. Những bài đánh giá có thể cho doanh nghiệp biết liệu rằng họ có theo kịp kỳ vọng của khách hàng hay không, điều này rất quan trọng để phát triển các chiến lược tiếp thị dựa trên mong muốn của khách hàng.

Nghiên cứu này được chia thành 5 phần, trong phần tiếp theo, nghiên cứu sẽ trình bày lý thuyết và các nghiên cứu liên quan trong lĩnh vực khai thác ý kiến người dùng, các phương pháp học máy để làm tiền đề cho mô hình và phương pháp nghiên cứu mà nghiên cứu này đề xuất. Sau đó, nghiên cứu sẽ tiến hành thực nghiệm và đánh giá kết quả trên một số mô hình học máy. Cuối cùng, kết luận và hướng phát triển sẽ được trình bày để tổng kết và nhìn nhận lại những ưu điểm, nhược điểm, cũng như hướng phát triển mà nghiên cứu này đã và sẽ thực hiện.

CƠ SỞ LÝ THUYẾT VÀ TÌNH HÌNH NGHIÊN CỨU LIÊN QUAN

Phần này tập trung khảo sát các nghiên cứu liên quan trong lĩnh vực khai thác ý kiến người dùng, phân tích quan điểm; đặc biệt là trong lĩnh vực dịch vụ trực tuyến. Các cách tiếp cận theo hướng học máy và từ vựng trong một số nghiên cứu cũng được khảo sát và phân tích để làm cơ sở cho bài nghiên cứu này.

Trí tuệ nhân tạo và Máy học

Trí tuệ nhân tạo (AI – Artificial Intelligence) để cập đến việc mô phỏng trí thông minh của con người trong máy móc, giúp máy móc có suy nghĩ giống như con người và bắt chước hành động của con người[11]. Ở dạng đơn giản nhất, trí tuệ nhân tạo là một lĩnh vực kết hợp giữa khoa học máy tính và xử lý dữ liệu nhằm giúp giải quyết vấn đề dựa trên máy móc. Các lĩnh vực hẹp như học máy và học sâu, thường được để cập đến cùng với trí tuệ nhân tạo. Các bộ môn này bao gồm các thuật toán trí tuệ nhân tạo nhằm tạo ra các hệ thống chuyên gia đưa ra các dự đoán hoặc phân loại dựa trên dữ liệu đầu vào.

Máy học (ML - Machine Learning) là một nhánh của trí tuệ nhân tạo và khoa học máy tính[12], tập trung vào việc sử dụng dữ liệu và thuật toán để bắt chước cách con người học, dần dần cải thiện độ chính xác. Trong nhiều thập kỷ qua, học máy là một thành phần quan trọng của lĩnh vực khoa học dữ liệu. Thông qua việc sử dụng các phương pháp thống kê, các thuật toán được đào tạo để đưa ra các phân loại hoặc dự đoán nhằm khám phá những hiểu biết quan trọng trong các dự án khai thác dữ liệu. Từ đó, những thông tin chú vị được phát hiện, sau đó thúc đẩy việc đưa ra quyết định trong các ứng dụng và doanh nghiệp.

Xử lý ngôn ngữ tự nhiên (NLP – Natural Language Processing) là một lĩnh vực liên ngành của ngôn ngữ học, khoa học máy tính và trí tuệ nhân tạo liên quan đến các tương tác giữa máy tính và ngôn ngữ con người[13], đặc biệt là cách lập trình máy tính để xử lý và phân tích lượng lớn dữ liệu ngôn ngữ tự nhiên. Mục tiêu là một máy tính có khả năng "hiểu" nội dung của các tài liệu, bao gồm các sắc thái ngữ cảnh của ngôn ngữ bên trong chúng. Sau đó, công nghệ này có thể trích xuất chính xác thông tin trong các tài liệu cũng như tự phân loại và sắp xếp các tài liệu. Những thách thức trong xử lý ngôn ngữ tự nhiên thường liên quan đến nhận dạng giọng nói, hiểu ngôn ngữ tự nhiên và tạo ngôn ngữ tự nhiên. Trong bối cảnh nghiên cứu này, các bình luận khách hàng cùng cấp được xem như ngôn ngữ tự nhiên và chúng ta cần phân tích để rút trích thông tin cần thiết, hữu ích.

Phân tích quan điểm và hành vi khách hàng

Phân tích quan điểm và hành vi khách hàng là việc thu thập và xử lý dữ liệu khách hàng để hiểu rõ hơn về trải nghiệm và quan điểm của khách hàng với sản phẩm hoặc dịch vụ mà họ đã sử dụng[14]. Phân tích hành vi

khách hàng cung cấp thông tin chi tiết có giá trị cho phép các thương hiệu đưa ra quyết định thông minh dựa trên dữ liệu, có khả năng cải thiện trải nghiệm mua sắm. Ngoài ra, thông tin thu thập được có thể giúp tạo các chiến dịch tiếp thị được tối ưu hóa. Trải nghiệm khách hàng được cải thiện thường dẫn đến tình cảm của khách hàng, doanh số bán hàng và thu nhập tốt hơn [15]. Quan điểm khi mua hàng là một trong các yếu tố quan trọng nhất giúp thu hút khách hàng mới và gia tăng lòng trung thành của khách hàng cũ.

Trong kinh doanh trực tuyến, khách hàng đôi khi tự thực hiện toàn bộ quá trình mua hàng mà không cần hỗ trợ dịch vụ khách hàng như kinh doanh trực tiếp [16,17]. Chẳng hạn, trong ngành khách sạn, các nền tảng trực tuyến cung cấp các dịch vụ bao gồm thông tin khách sạn và chức năng đặt phòng giúp khách hàng có thể dễ dàng thực hiện đặt phòng khách sạn bất cứ lúc nào [18]. Ngày nay, người tiêu dùng đang ngày càng truy cập các trang web trực tuyến nơi họ có thể truy cập nhiều thông tin liên quan về sản phẩm và dịch vụ trực tuyến [17]. Nhưng với sự gia tăng chóng mặt của các nền tảng công nghệ với các vấn đề bảo mật, người tiêu dùng chưa có niềm tin hoàn toàn với những thông tin về sản phẩm, dịch vụ trực tuyến. Để gia tăng niềm tin của khách hàng đối với sản phẩm dịch vụ, nhiều nền tảng trực tuyến cung cấp đánh giá của người tiêu dùng để cho phép người tiêu dùng thu thập thêm thông tin cho việc ra quyết định [19]. Người tiêu dùng hiện nay có xu hướng ủng hộ thông tin đến từ những người đi trước hơn là từ công ty cung cấp [17]. Và trước khi đưa ra quyết định, người tiêu dùng thường duyệt nhiều loại thông tin [20,21]. Trong số đó, các đánh giá trực tuyến là một nguồn thông tin quan trọng giúp họ có thể lựa chọn sản phẩm, dịch vụ đáp ứng tốt nhất nhu cầu, sở thích và giảm thiểu được rủi ro sử dụng [17,21].

Sự hài lòng của khách hàng là một dấu hiệu cho thấy niềm tin, quan điểm tích cực của khách hàng đối với một sản phẩm, dịch vụ [22]. Sự hài lòng có liên quan chặt chẽ đến thái độ và ý định của khách hàng, là một phần của hành vi khách hàng [23] và trực tiếp ảnh hưởng đến ý định hành vi khách hàng. Sự hài lòng là một yếu tố quan trọng ảnh hưởng đến lòng trung thành của khách hàng trực tuyến. Điều này đặt ra yêu cầu xây dựng lòng tin mạnh mẽ của người tiêu dùng, để người tiêu dùng quyết định làm trở nên dễ dàng hơn [24]. Và tái sử dụng dịch vụ chính là hành vi rõ nhất thể hiện sự hài lòng khách hàng. Tái sử dụng dịch vụ là một trong những yếu tố quan trọng nhất quyết định sự thành công của ngành công nghiệp đó. Bên trong đặc biệt là lĩnh vực du lịch, hiểu biết và dự đoán ý định quay lại của khách hàng và sử dụng lại các dịch vụ là một trong những vấn đề được chú ý nhất [18]. Bởi một khách hàng trong du lịch được phân thành khách truy cập lần đầu hoặc khách truy cập lại. Do đó, giữ chân khách hàng là một trong những vấn đề nghiên cứu được công nhận nhất trong ngành dịch vụ, đặc biệt là đối với kinh doanh trực tuyến.

Nghiên cứu của Lee và cộng sự [18] điều tra hành vi quay lại khách sạn của khách hàng bằng cách sử dụng quy mô lớn dữ liệu đánh giá của khách hàng. Nghiên cứu phân tích dữ liệu của 105.126 khách hàng của dịch vụ đặt phòng khách sạn trực tuyến, và tiến hành phân tích tình cảm về đánh giá phản hồi của người dùng. Qua so sánh khách truy cập một lần và khách truy cập lại, nghiên cứu cho thấy rằng đánh giá phản hồi của khách truy cập lại chứa nhiều từ hơn trong câu và hành vi quay trở lại khách sạn trong tương lai bộc lộ nhiều tình cảm tích cực/tiêu cực hơn so với du khách một lần. Mặt khác, các đánh giá phản hồi của một lần khách truy cập có xu hướng bao gồm nhiều từ phân tích và lo lắng hơn những từ của khách truy cập lại. Những phát hiện này có thể đóng vai trò là nền tảng cho việc sử dụng các phân tích dữ liệu trong nghiên cứu khách sạn và du lịch.

Nghiên cứu của Nobar và Rostamzadeh [25] đã điều tra tác động của sự hài lòng, trải nghiệm và lòng trung thành của khách hàng đối với sức mạnh thương hiệu trong ngành khách sạn. Nghiên cứu này sử dụng phương pháp tương quan, các nhóm được chọn là khách hàng của Pars Hotels. Số lượng mẫu là 384, dựa trên bảng lấy mẫu của Krejcie và Morgan. Kết quả nghiên cứu cho thấy kỳ vọng của khách hàng có ảnh hưởng nhiều nhất đến sự hài lòng của khách hàng với hệ số đường dẫn là 0,74. Mặt khác, lòng trung thành của khách hàng, với hệ số đường dẫn là 0,65, được coi là một yếu tố có ảnh hưởng. Nghiên cứu này giúp hiểu rằng sự hài lòng của khách hàng và mong đợi của khách hàng là những động lực tích cực cho lòng trung thành của khách hàng. Lòng trung thành của khách hàng cũng là một yếu tố dự báo mạnh mẽ về sức mạnh thương hiệu trong ngành khách sạn và du lịch.

Tại Việt Nam, nghiên cứu của Phụng và cộng sự [26] đã tiến hành thu thập dữ liệu từ nền tảng trực tuyến Agoda với 15.480 bình luận của khách hàng đối với các khách sạn tại Việt Nam. Nghiên cứu phân loại ý kiến khách hàng theo quan điểm "tiêu cực" hoặc "tích cực". Từ đó, đưa ra hướng giải quyết các bài toán về xếp hạng dịch vụ du lịch, đánh giá và cải tiến chất lượng dịch vụ, hệ thống gợi ý lựa chọn khách sạn du lịch. Kết quả nghiên cứu cho thấy thuật toán Logistic Regression (LR) và Support Vector Machine (SVM) là tốt nhất.

Nghiên cứu của Thảo và Thành [27] trong ngành thực phẩm thuần chay phân tích quan điểm theo 3 loại mức

độ: tích cực, tiêu cực và trung lập. Bộ dữ liệu bao gồm 17.892 đánh giá của khách hàng. Kết quả nghiên cứu cho thấy 20,8% quan điểm của khách hàng khi sử dụng sản phẩm chay là tích cực, 0,7% tiêu cực và 78,5% trung lập. Nghiên cứu này chứng tỏ rằng quan điểm của khách hàng đối với các khía cạnh về sản phẩm thuần chay theo nhận xét văn bản, khía cạnh chất lượng có ý nghĩa quan trọng hơn so với các khía cạnh khác. Vì thế doanh nghiệp có thể hiểu các yếu tố quan trọng để giải quyết những khó khăn trong quá trình kinh doanh.

Ứng dụng máy học vào phân tích quan điểm và hành vi khách hàng

Học máy ngày càng quan trọng trong thời đại số hiện nay, một lượng dữ liệu khổng lồ đang được tạo ra bởi khách hàng trên toàn cầu mỗi ngày. Các thuật toán máy học đã được sử dụng rộng rãi để phân tích quan điểm[28]. Một khách hàng tạo ra dữ liệu ở hầu hết mọi thời điểm trong ngày. Với việc khách hàng tạo ra lượng nguồn dữ liệu dồi dào, các doanh nghiệp đã bắt đầu tận dụng nó để đưa ra các quyết định dựa trên dữ liệu cho tổ chức của họ[29]. Tuy nhiên, rất khó để có thể đọc hết toàn bộ dữ liệu đó theo cách thủ công và đưa ra quyết định dựa trên những gì đọc được. Vì vậy, cần có một hệ thống để tự động phát hiện quan điểm của khách hàng từ nhận xét của họ để tạo thuận lợi cho quá trình ra quyết định[30]. Trước đây, sự phát triển của một thương hiệu chủ yếu phụ thuộc vào quảng cáo. Nhưng với sự tiến bộ của công nghệ để thu thập và xử lý lượng dữ liệu khổng lồ này, việc hiểu khách hàng của bạn để tiếp thị các sản phẩm đó đã trở thành chìa khóa để hiểu khách hàng nghĩ gì về sản phẩm, hiểu hành vi của người dùng hoặc nâng cao trải nghiệm của khách hàng. Quá trình thu thập và phân tích dữ liệu khách hàng để hiểu khách hàng được gọi là phân tích khách hàng. Nhiều nghiên cứu cho thấy phương pháp học máy được sử dụng để phân tích tình cảm, quan điểm của khách hàng[31–33]. Nghiên cứu của Yi và Liu[34] đã xây dựng hệ thống máy học phân loại quan điểm kết hợp với hệ khuyến nghị dựa trên luật kết hợp để tự động gợi ý khách hàng sản phẩm phù hợp dựa trên đánh giá với độ chính xác cao. Tác giả nghiên cứu[35] đã xây dựng mô hình phân tích quan điểm khách hàng cho các đánh giá sản phẩm với độ chính xác 81,75% từ thuật toán SVM sau khi qua bước xác thực chéo. Tại Việt Nam, nghiên cứu của Quỳnh và cộng sự[36] đã xây dựng mô hình phân tích quan điểm ở lĩnh vực thương mại điện tử với độ chính xác cao nhất lên đến 93%. Bằng cách thu thập đánh giá của khách hàng trên thị trường, trang web, cuộc gọi dịch vụ khách hàng hoặc khảo sát phản hồi của khách hàng, doanh nghiệp thu được một lượng lớn dữ liệu văn bản. Doanh nghiệp có thể sử dụng các phương pháp xử lý ngôn ngữ tự nhiên để xác định những khía cạnh nào của sản phẩm mà khách hàng đang nói đến và nêu quan điểm là tích cực, hay tiêu cực. Tiếng nói của khách hàng là rất quan trọng vì hiểu khách hàng sẽ giúp điều chỉnh sản phẩm hoặc dịch vụ phù hợp hơn, từ đó, nâng cao nhận thức về thương hiệu giữa các đối thủ cạnh tranh.

PHƯƠNG PHÁP, MÔ HÌNH ĐỀ XUẤT VÀ THỰC NGHIỆM

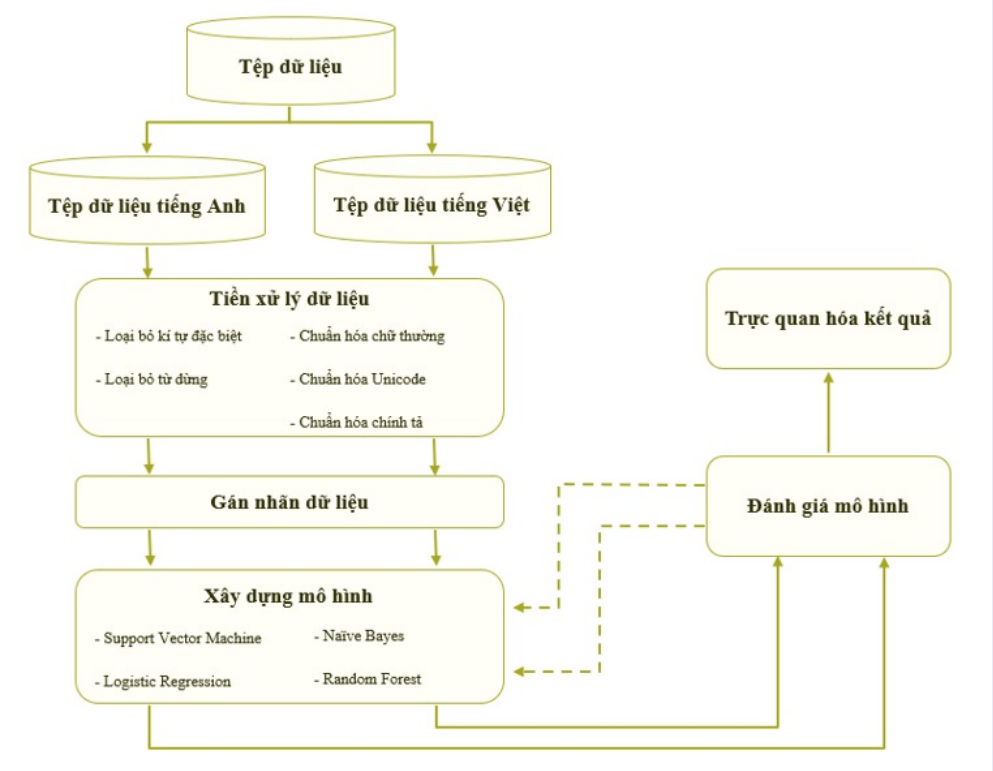
Trong phần này, nghiên cứu sẽ trình bày phương pháp thực hiện. Bên cạnh đó, mô hình mà nghiên cứu đề xuất cũng sẽ được giới thiệu, bao gồm khung giải pháp và thực nghiệm. Các bước chính của mô hình như tiền xử lý dữ liệu, huấn luyện, đánh giá mô hình sẽ được thể hiện chi tiết bên dưới. Phân tích quan điểm được xem là nhiệm vụ quan trọng trong phân tích cảm xúc. Trong phạm vi nghiên cứu này, nghiên cứu sẽ tập trung đề xuất mô hình phân tích quan điểm và hành vi của khách hàng dựa trên những đánh giá và bình luận thu thập được từ nền tảng Agoda.

Phương pháp nghiên cứu

Nghiên cứu này sử dụng phương pháp từ vựng và phương pháp học máy có giám sát nhằm phân tích cảm xúc và hành vi của người dùng trực tuyến trong lĩnh vực du lịch tại thị trường Việt Nam. Trong đó, phương pháp từ vựng dùng để trích xuất đặc trưng của từ, cụm từ từ một câu, một đoạn văn. Trong nghiên cứu này, phương pháp tách từ chủ yếu dựa trên phương pháp từ điển. Dựa trên nền tảng của phương pháp từ vựng, nghiên cứu áp dụng phương pháp học máy có giám sát với tập từ khóa chính là các từ vựng đã được trích xuất nhằm xác định mô hình học máy tối ưu giúp phân loại cảm xúc. Nghiên cứu tiến hành thực nghiệm trên cả tiếng Việt và tiếng Anh. Ngoài ra, nghiên cứu cũng áp dụng các phương pháp khác như phân tích và tổng hợp lý thuyết để xác định được khoảng trống nghiên cứu khoa học; phương pháp thu thập số liệu để thu thập dữ liệu hỗ trợ cho quá trình thực nghiệm; phương pháp thực nghiệm để đánh giá mô hình tối ưu.

Mô hình đề xuất

Hình 1 trình bày tổng quan về mô hình nghiên cứu. Kho dữ liệu sử dụng trong nghiên cứu được thu thập từ nền tảng đặt phòng trực tuyến Agoda. Dữ liệu thô sau khi được thu thập sẽ trải qua quá trình tiền xử lý dữ liệu để đảm bảo cho việc xây dựng mô hình có kết quả chính xác. Tiếp theo, nghiên cứu tiến hành



Hình 1: Mô hình nghiên cứu tổng thể (Nguồn: Tác giả đề xuất)

gán nhãn cho dữ liệu. Sau đó, nghiên cứu sẽ xây dựng mô hình máy học để phân tích quan điểm, nghiên cứu cũng sẽ thực hiện đánh giá độ chính xác của mô hình. Các bước này sẽ được thực hiện lặp lại liên tục để đạt được hiệu suất mô hình cao nhất. Cuối cùng, các kết quả từ mô hình sẽ được trực quan hóa bằng các biểu đồ.

Thực nghiệm mô hình

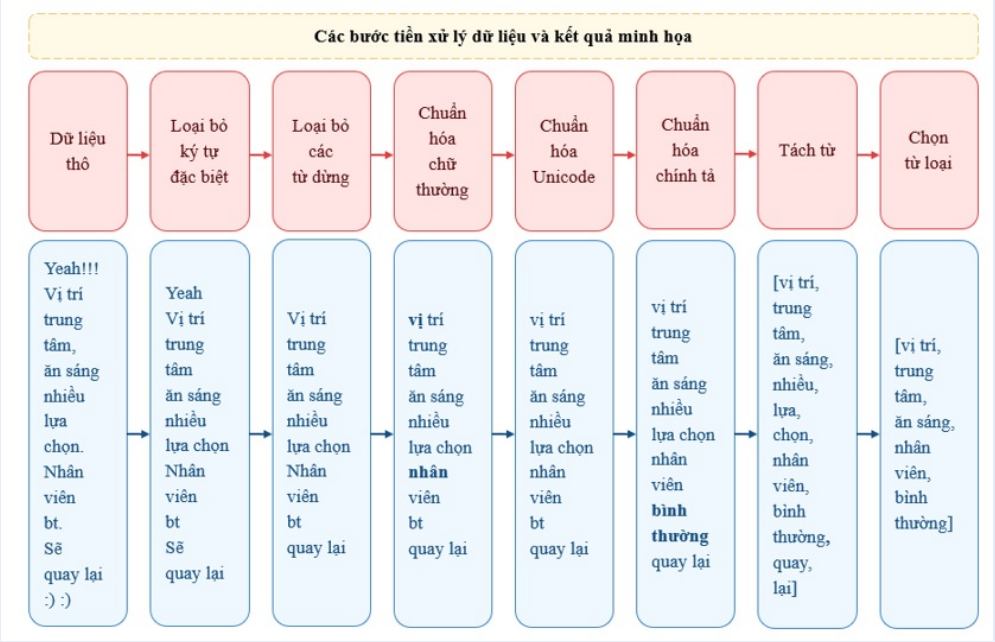
Thu thập dữ liệu

Để thu thập dữ liệu đánh giá từ nền tảng Agoda, nghiên cứu đã sử dụng 2 thư viện Python là Aiohttp và TheBeautifulSoup. Các yêu cầu HTTPs được gửi tới API của trang web để thu thập dữ liệu trực tiếp lưu trữ ở dạng JSON. Sau đó, từ dữ liệu thu thập được nghiên cứu sẽ chuyển đổi, trích xuất những đặc trưng quan trọng như "review_comments", "rating" và lưu trữ vào cơ sở dữ liệu.

Tiền xử lý

Đối với các mô hình máy học, độ sạch của dữ liệu là một yếu tố quan trọng ảnh hưởng lớn đến với hiệu suất mô hình[37]. Với tập dữ liệu thô thu thập được có thể chứa rất nhiều dữ liệu không liên quan hoặc vô nghĩa có thể ảnh hưởng xấu đến kết quả. Vì vậy, nghiên cứu sẽ làm sạch dữ liệu, quy trình thực hiện tuân theo sơ đồ ở Hình 2.

- Loại bỏ kí tự đặc biệt: Là những ký tự không thuộc bảng chữ cái hoặc ở dạng số, là những siêu liên kết, thẻ, khoảng trắng và dấu câu.
- Loại bỏ các từ dừng bao gồm mạo từ, đại từ, liên từ, giới từ và những từ cần thiết để hình thành câu nhưng không ảnh hưởng đến giá trị quan điểm cuối cùng của bình luận[38].
- Đổi về chữ viết thường để tránh sự khác biệt về chữ thường chữ hoa giữa các từ có cùng ý nghĩa.
- Chuyển đổi ký tự Unicode tổ hợp thành Unicode dựng sẵn vì những ký tự này tuy giống nhau nhưng với máy tính lại có bộ mã hoàn toàn khác nhau.
- Chuẩn hóa chính tả hoặc chuyển đổi các từ viết tắt thành một hình thức chuẩn.
- Tách từ: Nghiên cứu sử dụng thư viện "underthesea" cho tập dữ liệu tiếng Việt và thư viện "nltk" cho tập dữ liệu tiếng Anh để tiến hành tách từ.



Hình 2: Quy trình tiền xử lý dữ liệu (Nguồn: Kết quả thực nghiệm)

- Chọn từ loại: Nghiên cứu tiến hành trích xuất lại các từ loại là danh từ và tính từ để tiến hành huấn luyện nhằm tăng độ chính xác của mô hình.

Nghiên cứu hiện tại xem xét các từ khóa với trọng số như nhau, nghiên cứu chưa tập trung vào đánh giá trọng số và tần suất xuất hiện của từ khóa trong mỗi bình luận. Đây có thể được xem xét là một yếu điểm trong nghiên cứu làm cho độ chính xác của mô hình giảm xuống. Các nghiên cứu tiếp theo sẽ được cải tiến yếu điểm này.

Gán nhãn dữ liệu dựa vào điểm số đánh giá của khách hàng

Nghiên cứu này để cập đến việc phân tích ý kiến thông qua các kỹ thuật học máy có giám sát. Điều này có nghĩa là tập dữ liệu được gán nhãn đã chứa câu trả lời đúng. Sau khi huấn luyện và đánh giá kết quả, ta thu được một mô hình được sử dụng để phân loại quan điểm trong các bài đánh giá khách sạn mới chưa được gán nhãn. Sau khi thực hiện theo thứ tự và đầy đủ các bước tiền xử lý dữ liệu ở trên, nghiên cứu đã có được một tập dữ liệu sạch để chuẩn bị cho giai đoạn huấn luyện. Tiếp theo, nghiên cứu chia dữ liệu theo tỷ lệ 80:20 bằng phương pháp Hold-Out để lấy dữ liệu huấn luyện và dữ liệu đánh giá mô hình, tỷ lệ này được thực hiện trên cả tập dữ liệu tiếng Việt và tiếng Anh. Trong nghiên cứu này, nhãn dữ liệu được đánh dựa trên điểm số (rating) của khách hàng thông qua các bình luận trực tuyến. Các tác giả nghiên cứu[25] đã phân tích tác động của sự hài lòng, trải nghiệm của khách hàng và sự trung thành của khách hàng đối với thương hiệu dựa trên dữ liệu của ngành khách sạn. Kết quả từ nghiên cứu[25] cho thấy rằng, giá trị 7,03 được xem xét là ngưỡng để phân loại bình luận tích cực và tiêu cực. Bên cạnh đó, kế thừa từ nghiên cứu[26], các tác giả trong nghiên cứu này đã thực nghiệm các mô hình học máy trên tập dữ liệu Agoda và chọn ngưỡng 7,0 để phân loại "tích cực" và "tiêu cực". Kết quả từ mô hình trong nghiên cứu này cũng có độ chính xác cao. Ngoài ra, trong quá trình thực nghiệm, nghiên cứu này cũng tiến hành đánh giá dữ liệu thu thập được để tìm ra giá trị ngưỡng phù hợp. Nghiên cứu cũng nhận thấy rằng, các bình luận có điểm đánh giá nhỏ hơn 7,0 mang ý nghĩa tiêu cực (Negative), và ngược lại, các bình luận có điểm đánh giá lớn hơn 7,0 mang ý nghĩa tích cực (Positive)[25]. Do đó, nghiên cứu tiến hành lựa chọn giá trị 7,0 để làm ngưỡng phân loại tính chất của các bình luận. Bảng 1 là kết quả minh họa kết quả dữ liệu được gán nhãn.

Huấn luyện mô hình và Đánh giá mô hình

Để có được mô hình phù hợp với dữ liệu nghiên cứu, nghiên cứu đã sử dụng bốn thuật toán máy học như sau:

Logistic Regression: Là một thuật toán phân loại học máy được sử dụng để dự đoán xác suất của một biến

Bảng 1: Kết quả gán nhãn dữ liệu (Nguồn: Kết quả thực nghiệm)

Số thứ tự	Dữ liệu	Nhãn
1	dịch vụ bể bơi phải dùng nhờ ở khách sạn cạnh đó theo hướng dẫn của lễ tân	Tiêu cực
2	cơ sở vật chất cũ kỹ phòng hơi nhỏ nhân viên thiếu thân thiện	Tiêu cực
3	tất tuyệt vị trí cực kỳ trung tâm ngay sát phố đi bộ nguyễn huệ khách sạn mới sửa lại nên mới và rất tiện nghi mình thích cái đệm êm vô cùng.	Tích cực
4	giá tốt vị trí thuận lợi phòng sạch nhân viên thân thiện	Tích cực

phụ thuộc phân loại. Logistic Regression sử dụng hàm mất mát phức tạp được gọi là "hàm Sigmoid". Giả thuyết về Logistic Regression có xu hướng chặn hàm chi phí trong khoảng từ 0 đến 1. Do đó, Logistic Regression có thể được sử dụng trong phân tích quan điểm nhằm phân loại các đánh giá là tích cực hoặc là tiêu cực. Như kết quả được đề cập trong nghiên cứu[39], Logistic Regression có hiệu suất vượt trội, là một trong những thuật toán tốt nhất cho phân tích quan điểm.

Support Vector Machine: Là một phương pháp phân loại thống kê dựa trên việc tối đa hóa ranh giới giữa các điểm dữ liệu và siêu mặt phẳng phân tách[40]. Về cơ bản, thuật toán này sẽ định vị các biên tốt nhất có thể để phân tách giữa các dữ liệu tích cực cũng như tiêu cực và được sử dụng rộng rãi vì hiệu suất vượt trội của phương pháp Support Vector Machine so với các phương pháp khác được sử dụng trong hầu hết các mô hình máy học[41]. Bài nghiên cứu sử dụng phương pháp LinearSVC giúp cho Support Vector Machine linh hoạt hơn trong việc xác định giá trị mất mát để dự đoán độ chính xác của dữ liệu phân loại[42].

Naïve Bayes: Là một bộ phân loại đa thức dựa trên định lý Bayes, được sử dụng để phân loại quan điểm ở cấp độ tài liệu, kết quả và hiệu suất tương đối tốt[41]. Một đặc điểm của thuật toán Naïve Bayes là sự tồn tại của các biến đầu vào độc lập giả định sự hiện diện của một đối tượng đặc trưng từ một lớp độc lập với thành phần khác. Mô hình Naïve Bayes được chọn vì phương pháp này đã được triển khai rộng rãi trong phân tích tình cảm [32].

Random Forest: Là thuật toán máy học kết hợp được xây dựng từ các cây quyết định. Thuật toán này hoạt động dựa trên cách thức bỏ phiếu, có nghĩa là tất cả các cây quyết định trong mô hình sẽ thực hiện phân loại ra một lớp và lớp được bỏ phiếu nhiều nhất thì sẽ là đầu ra của mô hình[43]. Trong một khoảng thời gian rất ngắn kể từ khi được công bố, Random Forest là một những phương pháp phổ biến được sử dụng trong các mô hình phân tích quan điểm, phân tích dữ liệu[41].

Nhìn chung, có rất nhiều phương pháp học máy đã được ứng dụng trong bài toán phân tích quan điểm. Tuy nhiên, trong nghiên cứu này, sau quá trình khảo sát các nghiên cứu liên quan, nghiên cứu lựa chọn lại được 4 phương pháp chính như trên. Tiếp theo, quá trình thực nghiệm và đánh giá trên 4 phương pháp để lựa chọn mô hình tối ưu giúp cho giải pháp của nghiên cứu hoàn thiện và có độ tin cậy tốt hơn. Giờ đây, vấn đề đặt ra, làm thế nào để đánh giá và chọn ra các mô hình. Ngoài thuật toán học máy, sự thực thi của mô hình có thể phụ thuộc vào các yếu tố khác như sự phân bố của các lớp, chi phí phân loại sai, kích thước của tập huấn luyện và tập thử nghiệm, độ đo thực thi. Nghiên cứu này sẽ đánh giá tập trung vào khả năng dự đoán chính xác của mô hình hơn là tốc độ phân loại hay chi phí để xây dựng mô hình. Thông thường, hiệu quả của mô hình phân loại ý kiến được đánh giá dựa trên các chỉ số được mô tả ở Bảng 2, gồm có: Độ chính xác (Accuracy), Độ hội tụ (Precision), Độ bao phủ (Recall), và Giá trị trung bình điều hòa (F1_Score).

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

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

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

$$F1\_Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Khi xây dựng các mô hình học máy, một số vấn đề như mô hình quá khớp (Overfitting) hay mô hình dưới khớp (Underfitting) cần được quan tâm. Nghiên cứu này nhận thấy lượng dữ liệu hiện tại đủ lớn để có thể xây dựng được mô hình tổng quan. Ngoài ra, nghiên cứu sẽ sử dụng giá trị lỗi để xem xét lựa chọn mô hình phù hợp nhằm đảm bảo dữ liệu không bị quá khớp hoặc dưới khớp.

KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

Các kết quả thu thập dữ liệu, tiền xử lý, huấn luyện và đánh giá mô hình được trình bày ở phần này. Cùng với đó là những kết quả được trực quan hóa và các thảo luận liên quan đến chủ đề mà nghiên cứu thực hiện.

Bảng 2: Ma trận nhầm lẫn (Confusion matrix)

	Dự đoán: Tích cực (1)	Dựa đoán: Tiêu cực (0)
Thực tế: Tích cực (1)	TP	FN
Thực tế: Tiêu cực (0)	FP	TN

Bảng 3: Tập dữ liệu huấn luyện và đánh giá (Nguồn: Kết quả thực nghiệm)

Tập dữ liệu	Tiếng Việt	Tiếng Anh
Huấn luyện (80%)	48.174	170.094
Kiểm tra (20%)	12.044	42.523
Tổng (100%)	60.218	212.617

Tập dữ liệu

Dữ liệu được thu thập với 60.218 phản hồi tiếng Việt và 212.617 phản hồi tiếng Anh của khách hàng từ nền tảng Agoda. Mỗi dòng chứa những đặc trưng liên quan đến mỗi bình luận của khách hàng, gồm các thông tin như tên khách sạn, địa chỉ, tên khách hàng bình luận, thời gian bình luận, nội dung bình luận, điểm đánh giá của khách hàng đối với khách sạn đó. Tập dữ liệu này sẽ được đưa vào bước tiền xử lý, làm sạch để cung cấp đầu vào cho quá trình huấn luyện mô hình. Bảng 3 mô tả số lượng dữ liệu huấn luyện và kiểm tra trên cả 2 tập dữ liệu tiếng Việt và tiếng Anh. Bên cạnh đó, Bảng 4 và Bảng 5 lần lượt mô tả tập dữ liệu thu thập được từ nền tảng Agoda với ngôn ngữ tiếng Việt và tiếng Anh.

Tần suất xuất hiện của các từ mô tả sắc thái

Biểu đồ đám mây gồm các từ (words) hoặc cụm từ (phrases) có trọng số thường được sử dụng để mô tả siêu dữ liệu từ khóa trên các trang web hoặc để trực quan hóa văn bản dạng tự do. Hình 3 và Hình 4 lần lượt biểu diễn các từ thường xuyên xuất hiện trong bình luận của khách hàng trên tập dữ liệu tiếng Việt và tiếng Anh.

Các từ khóa tiêu cực và tích cực được biểu diễn bằng biểu đồ đám mây giúp nhà quản trị nắm bắt thông tin dễ dàng. Các từ thường là các từ đơn và tầm quan trọng của mỗi từ được thể hiện bằng kích thước hoặc màu sắc phông chữ. Các từ lớn hơn có nghĩa là trọng lượng lớn hơn hay nói cách khác là xuất hiện trong bình luận nhiều hơn. Đó cũng là những vấn đề khách hàng thường xuyên quan tâm và trao đổi.

Kết quả thực nghiệm và thảo luận

Trong nghiên cứu này, mô hình được huấn luyện thông qua các thuật toán: Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM). Kết quả huấn luyện thể hiện trong Bảng 6, Hình 5 và Hình 6 với tập dữ liệu tiếng Việt. Trong khi đó, Bảng 7, Hình 7 và Hình 8 trình bày kết quả đánh giá của các mô hình học máy trong tập dữ liệu tiếng Anh.

Kết quả huấn luyện cho thấy các mô hình RF, NB, LR và SVM có độ chính xác khá cao (lần lượt là 90%; 90%; 88%; và 83% ở tiếng Việt và 88%; 86%; 90%; và 90% ở tiếng Anh). Nghiên cứu cũng đã xem xét giá trị lỗi để đảm bảo các mô hình hiện tại không bị quá khớp hay dưới khớp. Tiếp theo, ở kết quả đường cong ROC của các mô hình tiếng Việt và tiếng Anh, bài nghiên cứu sử dụng thang đo lường trung bình vĩ mô, là kết quả trung bình cộng theo lớp thay vì trung bình vi mô của toàn bộ kết quả dự đoán. Đường cong ROC ở các mô hình cho thấy rằng SVM là mô hình tốt nhất cho việc phân tích quan điểm, tại 80% ở mô hình tiếng Việt và 73% ở mô hình tiếng Anh. Nghĩa là các mô hình này tương đối phù hợp với tập dữ liệu huấn luyện. Trong đó, có mô hình SVM và LR là tốt nhất. Do đó, các ứng dụng tiếp theo có thể dùng hai mô hình này như một công cụ để phân loại ý kiến cho các dữ liệu bình luận chưa được phân loại hoặc các dữ liệu bình luận mới phát sinh mà không cần phải huấn luyện lại. Kết quả nghiên cứu này đã giúp xác định phương pháp và công cụ phân loại ý kiến phù hợp. Đây được xem là bước quan trọng nhất của quy trình khai thác ý kiến, làm nền tảng cho việc ứng dụng khai thác ý kiến, phân tích quan điểm người dùng trong nhiều lĩnh vực.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Đóng góp của nghiên cứu

Với kho nội dung trực tuyến khổng lồ và các đặc điểm dữ liệu, các doanh nghiệp dần chuyển đổi công việc của một nhà tiếp thị kỹ thuật số từ một người kể chuyện kinh doanh thành một nhà quản lý công nghệ. Để hợp lý hóa các quy trình và tăng năng suất, các nhà tiếp thị kỹ thuật số (Cả hiện tại và tương lai) phải bắt đầu sử dụng các công cụ học máy để tự động hóa các

Bảng 4: Kết quả thu thập dữ liệu tiếng Việt (Nguồn: Kết quả thực nghiệm)

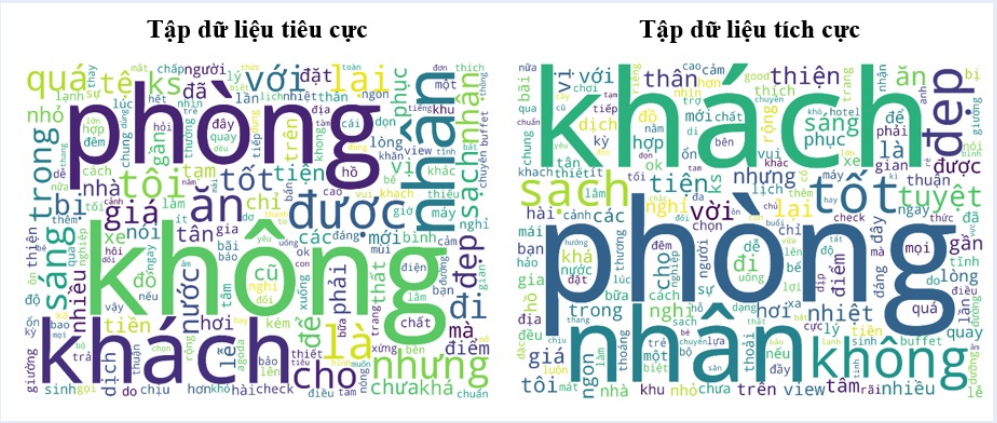
STT	Tỉnh, thành phố	Số lượng bình luận
1	Nha Trang	7.603
2	Vũng Tàu	6.712
3	Đảo Phú Quốc	5.377
4	Đà Nẵng	5.275
5	Hồ Chí Minh	4.781
6	Phan Thiết	4.258
7	Hà Nội	3.855
8	Hội An	3.446
9	Huế	2.797
10	Cần Thơ	2.262
11	Đà Lạt	1.975
12	Khác	11.877
Tổng cộng		60.218

Bảng 5: Kết quả thu thập dữ liệu tiếng Anh (Nguồn: Kết quả thực nghiệm)

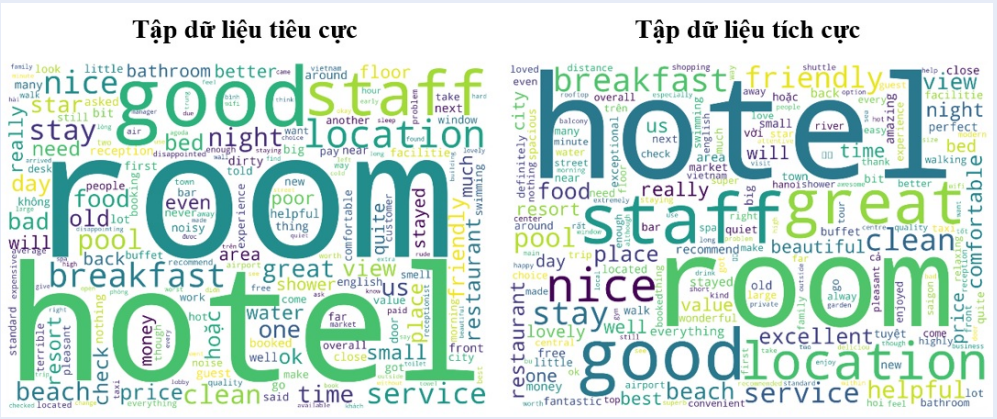
STT	Tỉnh, thành phố	Số lượng bình luận
1	Hồ Chí Minh	64.580
2	Hội An	22.953
3	Hà Nội	20.665
4	Đà Nẵng	20.637
5	Nha Trang	17.780
6	Đảo Phú Quốc	14.339
7	Huế	10.667
8	Phan Thiết	8.671
9	Vũng Tàu	7.837
10	Cần Thơ	3.784
11	Sapa	3.264
12	Khác	17.440
Tổng cộng		212.617

Bảng 6: Kết quả đánh giá mô hình tiếng Việt (Nguồn: Kết quả thực nghiệm)

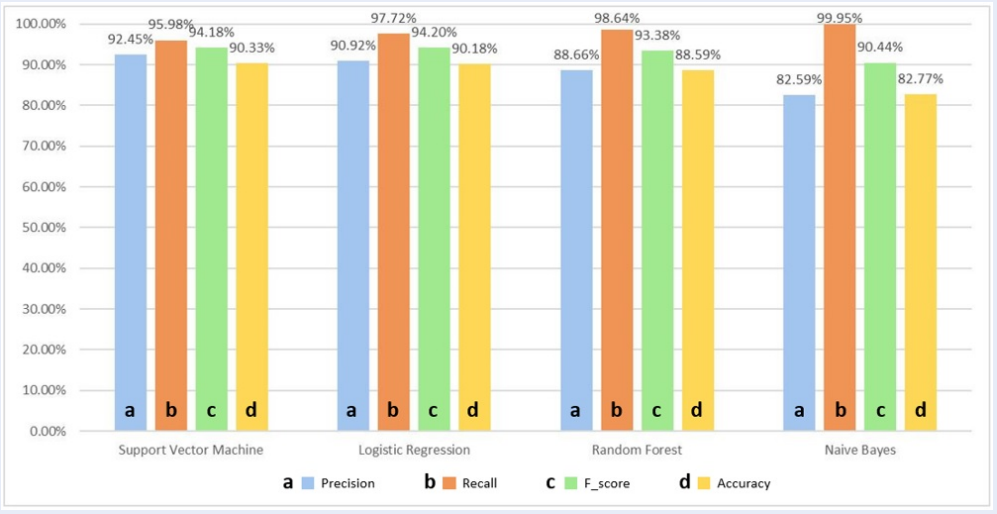
Thuật toán	Support Vector Machine	Logistic Regression	Random Forest	Naive Bayes
Precision	92,45%	90,92%	88,66%	82,59%
Recall	95,98%	97,72%	98,64%	99,95%
F_score	94,18%	94,20%	93,38%	90,44%
Accuracy	90,33%	90,18%	88,59%	82,77%



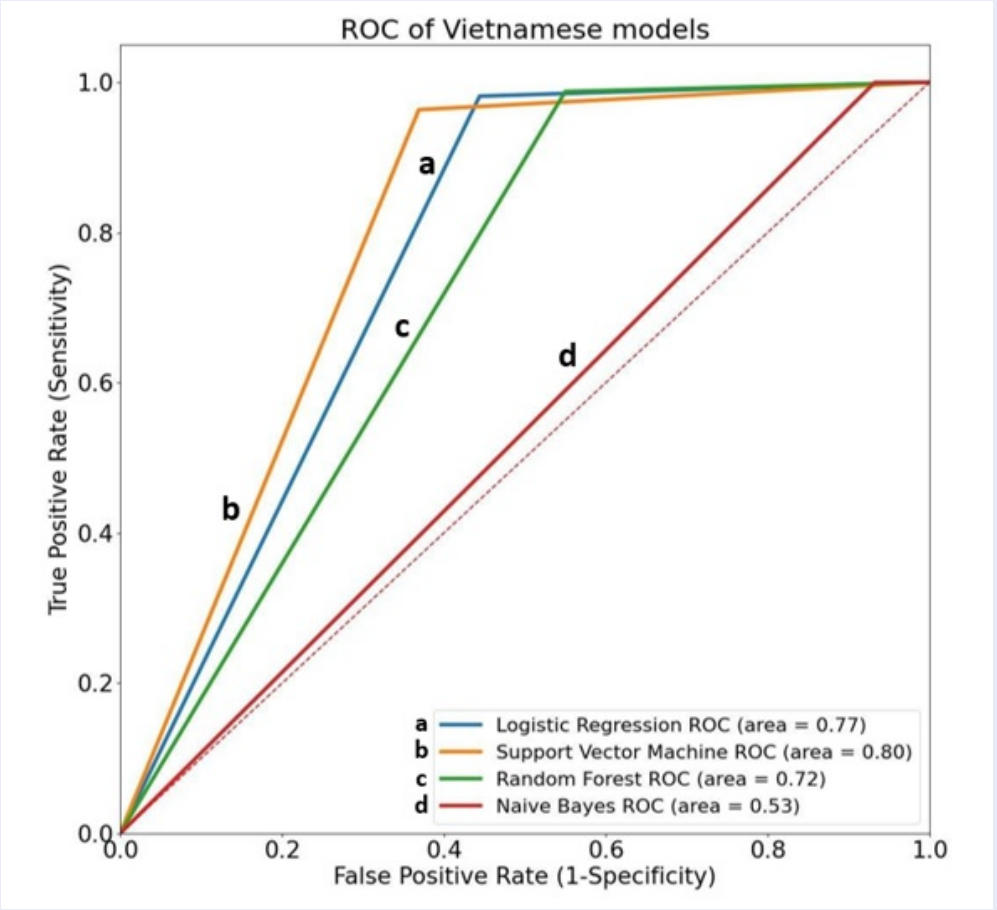
Hình 3: Đám mây từ tiếng Việt (Nguồn: Kết quả thực nghiệm)



Hình 4: Đám mây từ tiếng Anh (Nguồn: Kết quả thực nghiệm)



Hình 5: Kết quả đánh giá mô hình tiếng Việt (Nguồn: Kết quả thực nghiệm)



Hình 6: Đường cong ROC của các mô hình tiếng Việt (Nguồn: Kết quả thực nghiệm)

Bảng 7: Kết quả đánh giá mô hình tiếng Anh (Nguồn: Kết quả thực nghiệm)

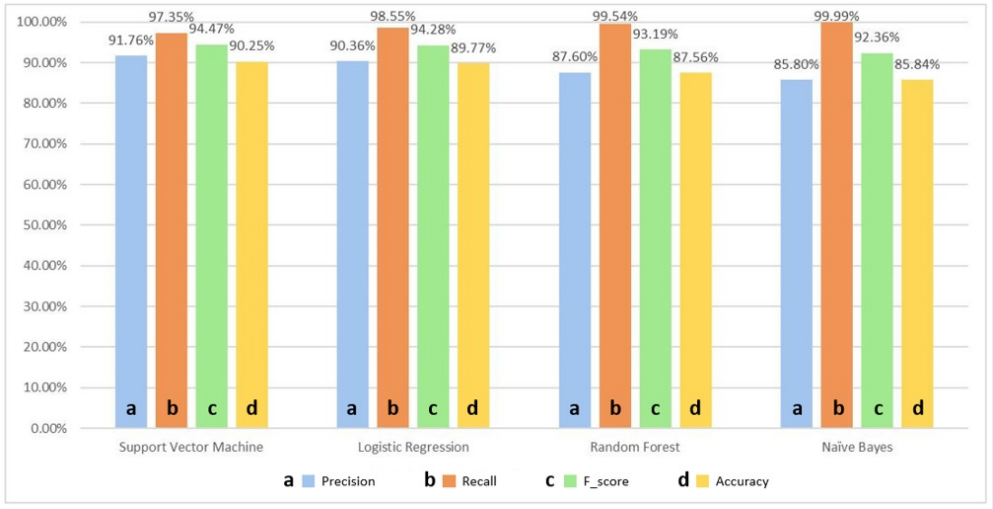
Thuật toán	Support Vector Machine	Logistic Regression	Random Forest	Naïve Bayes
Precision	91,76%	90,36%	87,60%	85,80%
Recall	97,35%	98,55%	99,54%	99,99%
F_score	94,47%	94,28%	93,19%	92,36%
Accuracy	90,25%	89,77%	87,56%	85,84%

quy trình và sử dụng dữ liệu hiệu quả nhất. Nếu mục tiêu của các nhà quản trị là tăng mức độ tương tác và nhận thức về thương hiệu với khách hàng tiềm năng, thì điều quan trọng là họ phải hiểu khách hàng của mình, đặc biệt là quan điểm quan điểm của họ. Kết quả nghiên cứu có một số đóng góp chính như sau:

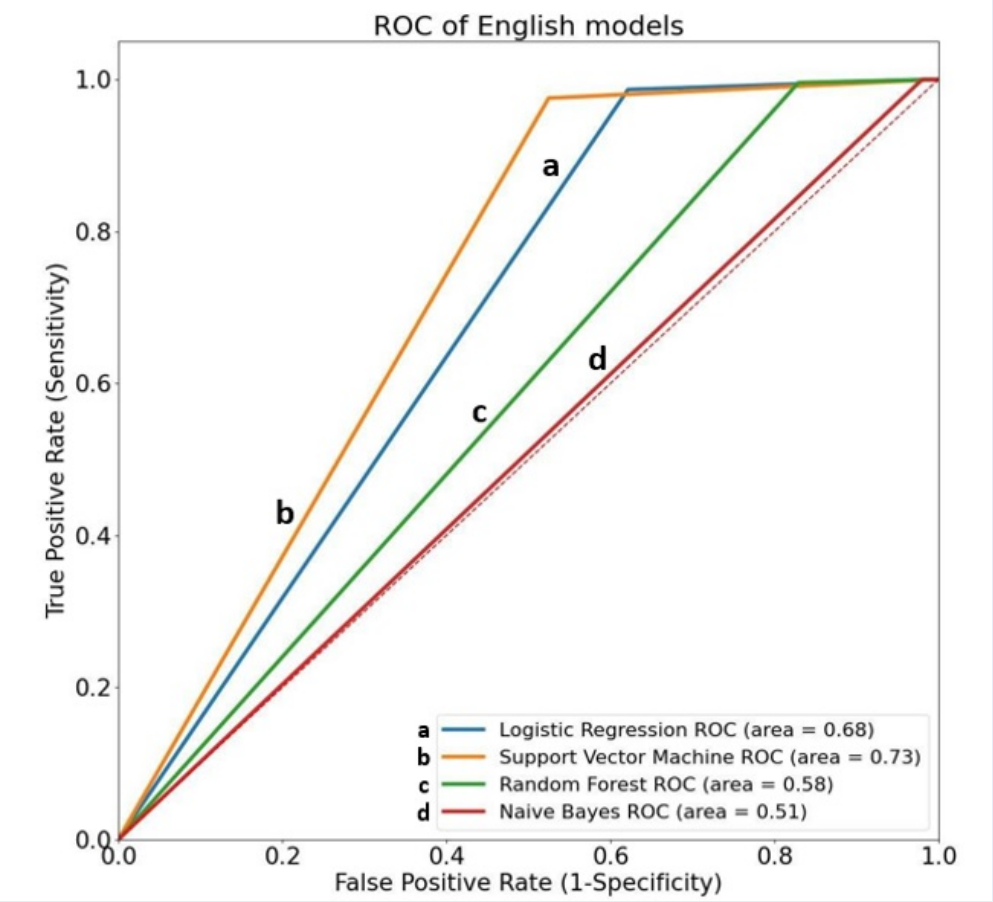
Thứ nhất, nghiên cứu thu thập bộ dữ liệu là các bình luận trực tuyến của người dùng trong lĩnh vực khách sạn. Bộ dữ liệu có kích thước 272.835 bình luận đủ lớn để xây dựng các mô hình máy học với hiệu suất cao.

Thứ hai, nghiên cứu đề xuất mô hình học máy có giám sát ứng dụng trong phân tích quan điểm người dùng trực tuyến. Kết quả nghiên cứu cho thấy mô hình có độ chính xác cao và phù hợp với yêu cầu thực tiễn của các doanh nghiệp hiện nay.

Cuối cùng, bài viết đóng góp một nghiên cứu liên ngành kết hợp các phương pháp thu thập và phân tích dữ liệu và phân tích quan điểm người dùng trong việc quản trị trải nghiệm, quan điểm và hành vi của khách hàng.



Hình 7: Kết quả đánh giá mô hình tiếng Anh (Nguồn: Kết quả thực nghiệm)



Hình 8: Đường cong ROC của các mô hình tiếng Anh (Nguồn: Kết quả thực nghiệm)

Hướng phát triển

Trong tương lai gần, trí tuệ nhân tạo hay học máy sẽ không thể thay thế các công việc tiếp thị kỹ thuật số. Thay vào đó, nó sẽ giúp mở rộng khả năng của nhà quản lý hiện đại, cung cấp cơ sở để hỗ trợ nhà quản trị ra quyết định, giúp việc phân tích trải nghiệm người dùng được tốt hơn. Trong các nghiên cứu tiếp theo, nghiên cứu sẽ mở rộng hệ thống theo hướng xử lý dữ liệu thu thập theo thời gian thực và ứng dụng vào các lĩnh vực kinh doanh, quản lý khác. Đối với quá trình tiền xử lý dữ liệu và gán nhãn, mô hình có thể được nâng cấp để giải quyết các lĩnh vực khó khăn hơn như thư rác và giả mạo, phủ định, châm biếm... Ngoài ra, những nghiên cứu trong tương lai có thể mở rộng đề tài thành phát hiện cảm xúc, một chủ đề sâu hơn của phân tích quan điểm nhằm hiểu hơn về hành vi của khách hàng. Tận dụng sức mạnh của trí tuệ nhân tạo và học máy để nâng cao năng lực của doanh nghiệp, nhận diện các vấn đề hiện tại, khắc phục các nhược điểm hiện có và đồng thời bắt đầu tạo ra tác động trong tương lai.

DANH SÁCH TỪ VIẾT TẮT

GDP: Gross Domestic Product – Tổng sản phẩm nội địa
SVM: Support Vector Machine
LR: Logistic Regression
RF: Random Forest
NB: Naïve Bayes
eWOM: electronic Word of Mouth – Tiếp thị truyền miệng
AI: Artificial Intelligence - Trí tuệ nhân tạo
ML: Machine Learning - Học máy
NLP: Natural Language Processing - Xử lý ngôn ngữ tự nhiên
API: Application Programming Interface – Giao diện lập trình ứng dụng
HTTPs: HyperText Transfer Protocol security – Giao thức truyền tin siêu văn bản

XUNG ĐỘT LỢI ÍCH

Nhóm tác giả xin cam đoan rằng không có bất kỳ xung đột lợi ích nào trong công bố bài báo.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Các tác giả cùng đóng góp trong quá trình lên ý tưởng và đề xuất mô hình.
Lê Bá Thiển chịu trách nhiệm thu thập dữ liệu.
Nguyễn Văn Hồ và Hồ Trung Thành tiến hành làm sạch dữ liệu, thực nghiệm trên tập dữ liệu tiếng Việt.
Lê Thị Kim Hiền và Nguyễn Quang Phúc tiến hành làm sạch dữ liệu, thực nghiệm trên tập dữ liệu tiếng Anh.
Hồ Trung Thành, Nguyễn Văn Hồ và Lê Hoành Sử chịu trách nhiệm lên bản thảo nghiên cứu.
Cuối cùng, Hồ Trung Thành và Lê Bá Thiển chịu trách nhiệm chỉnh sửa và trao đổi, phản hồi với người phản biện, nhà xuất bản.

TÀI TRỢ NGHIÊN CỨU

Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số DS2022-34-01.

TÀI LIỆU THAM KHẢO

1. Mishra A, Satish SM. eWOM: Extant Research Review and Future Research Avenues. The Journal of Decision Maker. 2016;41(3):222-233;Available from: https://doi.org/10.1177/0256090916650952.
2. Erkan I, Evans C. The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. Comput Human Behav 2016;61:47-55;Available from: https://doi.org/10.1016/J.CHB.2016.03.003.
3. Nusair KK, Bilgihan A, Okumus F, Cobanoglu C. Generation Y travelers' commitment to online social network websites. Tour Manag. 2013;35:13-22;Available from: https://doi.org/10.1016/j.tourman.2012.05.005.
4. Blomberg-Nygard A, Anderson CK. United Nations World Tourism Organization Study on Online Guest Reviews and Hotel Classification Systems: An Integrated Approach. Service Science. 2016;8:139-51;Available from: https://doi.org/10.1287/serv.2016.0139.
5. Berne-Manero C, Gómez-Campillo M, Marzo-Navarro M, Pedraja-Iglesias M. Reviewing the online tourism value chain. Adm Sci. 2018;8;Available from: https://doi.org/10.3390/admsci8030048.
6. Thu HNT. Measuring guest satisfaction from online reviews: Envidence in Vietnam. Cogent Soc Sci 2020;6;Available from: https://doi.org/10.1080/23311886.2020.1801117.
7. Mankad S, Han HS, Goh J, Gavirneni S. Understanding Online Hotel Reviews Through Automated Text Analysis. 2016;8:124-38;Available from: https://doi.org/10.1287/serv.2016.0126.
8. He W, Tian X, Tao R, Zhang W, Yan G, Akula V. Application of social media analytics: A case of analyzing online hotel reviews. Online Information Review. 2017;41:921-35;Available from: https://doi.org/10.1108/OIR-07-2016-0201.
9. Ghani NA, Hamid S, Targio Hashem IA, Ahmed E. Social media big data analytics: A survey. Comput Human Behav. 2019;101:417-28;Available from: https://doi.org/10.1016/J.CHB.2018.08.039.
10. Sarkar S, Palit S. Sentiment Analysis of Product Reviews of Ecommerce Websites. 2020:55-63;Available from: https://doi.org/10.1007/978-981-15-1059-5_7.
11. Brandon W. Jackson. Artificial Intelligence and the fog of innovation: A Deep-Dive on governance and the liability of autonomous systems, 35 Santa, Clara High Tech. L.J. 2019:35;Available from: https://digitalcommons.law.scu.edu/chtlj/vol35/iss4/1/.
12. Mehryar Mohri. Afshin Rostamizadeh, and Ameet Talwalkar, Foundations of Machine Learning, MIT Press, Second Edition. 2018;.
13. Chowdhary KR. Natural Language Processing. Fundamentals of Artificial Intelligence. 2022:603-649;Available from: https://doi.org/10.1007/978-81-322-3972-7_19.
14. Jain VK, Kumar S. Improving Customer Experience Using Sentiment Analysis in E-Commerce. 2017;Available from: https://doi.org/10.4018/978-1-5225-0997-4.ch012.
15. Manning H, Czarnecki D. Customer Experience Drives Revenue Growth, 2016 Business Case: The Customer Experience Ecosystem Playbook. 2016;.

16. Rita P, Oliveira T, Farisa A. The impact of e-service quality and customer satisfaction on customer behavior in online shopping. Heliyon 2019;5:e02690;Available from: https://doi.org/10.1016/J.HELIYON.2019.E02690.
17. Park E, Kang J, Choi D, Han J. Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews. 2018;23:605-11;Available from: https://doi.org/10.1080/13683500.2018.1549025.
18. Lee CKH, Tse YK, Zhang M, Ma J. Analysing online reviews to investigate customer behaviour in the sharing economy: The case of Airbnb. Information Technology and People 2020;33:945-61;Available from: https://doi.org/10.1108/ITP-10-2018-0475.
19. Kwok L, Xie KL. Factors contributing to the helpfulness of online hotel reviews: Does manager response play a role? International Journal of Contemporary Hospitality Management 2016;28:2156-77;Available from: https://doi.org/10.1108/IJCHM-03-2015-0107.
20. Moon S, Kim MY, Bergey PK. Estimating deception in consumer reviews based on extreme terms: Comparison analysis of open vs. closed hotel reservation platforms. J Bus Res 2019;102:83-96;Available from: https://doi.org/10.1016/J.JBUSRES.2019.05.016.
21. Zeng G, Cao X, Lin Z, Xiao SH. When online reviews meet virtual reality: Effects on consumer hotel booking. Ann Tour Res 2020;81:102860;Available from: https://doi.org/10.1016/J.ANNALS.2020.102860.
22. Udo GJ, Bagchi KK, Kirs PJ. An assessment of customers' e-service quality perception, satisfaction and intention. Int J Inf Manage 2010;30:481-92;Available from: https://doi.org/10.1016/J.IJINFOMGT.2010.03.005.
23. Holloway BB, Wang S, Parish JT. The role of cumulative online purchasing experience in service recovery management. Journal of Interactive Marketing 2005;19:54-66;Available from: https://doi.org/10.1002/DIR.20043.
24. Punyatoya P. Effects of cognitive and affective trust on online customer behavior. Marketing Intelligence and Planning 2019;37:80-96;Available from: https://doi.org/10.1108/MIP-02-2018-0058.
25. Nobar HBK, Rostamzadeh R. The impact of customer satisfaction, customer experience and customer loyalty on brand power: empirical evidence from hotel industry. Journal of Business Economics and Management. 2018;19:417-30;Available from: https://doi.org/10.3846/JBEM.2018.5678.
26. Kim Phụng T, Tế NA, Thị Thu Hà T. Tiếp cận phương pháp máy học trong khai thác ý kiến khách hàng trực tuyến. Tạp Chí Nghiên Cứu Kinh Tế và Kinh Doanh Châu Á. 2020;30:27-41;Available from: https://doi.org/10.46223/HCMCOUJS.econ.vi.16.2.612.2021.
27. Thi T, Thao H, Thanh HT. Exploring consumer opinions on vegetarian food by sentiment analysis method. HCMCOUJS-Economics and Business Administration. 2022;13(2);.
28. Agarwal B, Mittal N. Machine Learning Approach for Sentiment Analysis. 2016:21-45;Available from: https://doi.org/10.1007/978-3-319-25343-5_3.
29. Luo JM, Vu HQ, Li G, Law R. Understanding service attributes of robot hotels: A sentiment analysis of customer online reviews. Int J Hosp Manag. 2021;98:103032;Available from: https://doi.org/10.1016/J.IJHM.2021.103032.
30. Hasanli H, Rustamov S. Sentiment Analysis of Azerbaijani twits Using Logistic Regression, Naive Bayes and SVM. 13th IEEE International Conference on Application of Information and Communication Technologies, AICT 2019 - Proceedings 2019;Available from: https://doi.org/10.1109/AICT47866.2019.8981793.
31. Zvarevashe K, Olugbara OO. A framework for sentiment analysis with opinion mining of hotel reviews. 2018 Conference on Information Communications Technology and Society, ICTAS 2018 - Proceedings 2018:1-4;Available from: https://doi.org/10.1109/ICTAS.2018.8368746.
32. Laksono RA, Sungkono KR, Sarno R, Wahyuni CS. Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes. Proceedings of 2019 International Conference on Information and Communication Technology and Systems, ICTS 2019. 2019:49-54;Available from: https://doi.org/10.1109/ICTS.2019.8850982.
33. Park E, Kang J, Choi D, Han J. Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews. 2018;23:605-11;Available from: https://doi.org/10.1080/13683500.2018.1549025.
34. Yi S, Liu X. Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. Complex Intell. 2020;6:621-634;Available from: https://doi.org/10.1007/s40747-020-00155-2.
35. Singla Z, Randhawa S, Jain S. Sentiment analysis of customer product reviews using machine learning. Proceedings of 2017 International Conference on Intelligent Computing and Control, I2C2 2017 2018;2018-January:1-5;Available from: https://doi.org/10.1109/I2C2.2017.8321910.
36. Thuy Q, Thuy QNT, Bich NBN, Bao TNT, Nhat NT, Vo TB, et al. Customer experience discovery model based on sentiment analysis and machine learning method. VNUHCM Journal of Economics, Business and Law. 2022;6(3):3277-3290;Available from: https://doi.org/10.32508/stdjelm.v6i3.1030.
37. Chu X, Ilyas IF, Krishnan S, Wang J. Data cleaning: Overview and emerging challenges. Proceedings of the ACM SIGMOD International Conference on Management of Data. 2016;26-June-2016:2201-2206;Available from: https://doi.org/10.1145/2882903.2912574.
38. Nandal N, Tanwar R, Pruthi J. Machine learning based aspect level sentiment analysis for Amazon products. Spatial Information Research. 2020;28:601-607;Available from: https://doi.org/10.1007/s41324-020-00320-2.
39. Ahuja R, Chug A, Kohli S, Gupta S, Ahuja P. The Impact of Features Extraction on the Sentiment Analysis. Procedia Comput Sci. 2019;152:341-348;Available from: https://doi.org/10.1016/J.PROCS.2019.05.008.
40. Amrani Y, Lazaar M, el Kadirp KE. Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Procedia Comput Sci. 2018;127:511-520;Available from: https://doi.org/10.1016/J.PROCS.2018.01.150.
41. Shah Muhammad S, Awan S, Ahmad M, Aftab S, Ahmad S. Machine Learning Techniques for Sentiment Analysis: A Review. INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING. 2017;8;.
42. sklearn.svm.SVC - scikit-learn 1.2.0 documentation n.d. (accessed January 11, 2023);Available from: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.
43. Baid P, Gupta A, Chaplot N. Sentiment Analysis of Movie Reviews using Machine Learning Techniques. Article in International Journal of Computer Applications. 2017;179(7):975-8887;Available from: https://doi.org/10.5120/ijca2017916005.

Science & Technology Development Journal – Economics - Law and Management 2023, 7(1):4089-4103

Open Access Full Text Article

Research Article

Analysis of online user sentiment and behavior in the tourism sector in Vietnam based on reviews and comments

Thanh Ho[1,2], Van Ho Nguyen[1,2], Hoanh Su Le[1,2], Thi Kim Hien Le[1,2], Phuc Nguyen[1,2], Thien Le[1,2,*]

Use your smartphone to scan this QR code and download this article

ABSTRACT

As an attractive destination, Vietnam's tourism industry attracts a large number of tourists and has become a spearhead industry that plays an important role in the country's GDP. The large number of tourists leads to the development of the hotel industry, especially online hotel business services. Today, the explosion of social platforms and e-commerce sites has helped businesses collect large amounts of data from customer feedback. Exploiting this endless resource will give businesses a competitive advantage over competitors. Within the scope of this research, the sentiment analysis based on reviews and comments is focused on considering. Sentiment analysis is very useful to help business managers understand the needs and desires of customers. From there, come up with appropriate strategies to develop and improve the quality of products and services. The data of the study was collected on the online travel platform agoda.com with 272,835 comments. Research using machine learning models and algorithms Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB) to evaluate and analyze customer views in the business sector online hotel business in Vietnam. The research results have high accuracy up to 90% in both English and Vietnamese.

Key words: customer feedback, sentiment analysis, behavioral analysis, machine learning

[1]*University of Economics and Law, Ho Chi Minh City, Vietnam*

[2]*Vietnam National University, Ho Chi Minh City, Vietnam*

Correspondence

Thien Le, University of Economics and Law, Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City, Vietnam

Email: thienlb.ktl@uel.edu.vn

History

- Received: 02-02-2023
- Accepted: 24-03-2023
- Published: 31-3-2023

DOI : https://doi.org/10.32508/stdjelm.v7i1.1187

Check for updates

VNU-HCM Press

Cite this article : Ho T, Nguyen V H, Le H S, Le T K H, Nguyen P, Le T. **Analysis of online user sentiment and behavior in the tourism sector in Vietnam based on reviews and comments**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 2023, 7(1):4089-4103.