Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế-Luật và Quản lý, 4(4):1081-1092

Open Access Full Text Article

Bài nghiên cứu

Phân tích ý kiến khách hàng trực tuyến trong lĩnh vực khách sạn tiếp cận theo mô hình chủ đề

Nguyễn Văn Hồ[1], Hồ Trung Thành[2,*]

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

TÓM TẮT

Trong những năm gần đây, với sự phát triển của công nghệ và Internet, người dùng có thể dễ dàng đưa ra ý kiến đánh giá nhận xét của mình về các sản phẩm, dịch vụ của doanh nghiệp. Những thông tin này được lưu trữ dưới dạng dữ liệu văn bản, và là một nguồn dữ liệu khổng lồ để khai phá. Để tiếp tục phát triển đáp ứng nhu cầu người dùng, các doanh nghiệp cần biết những vấn đề khách hàng đang thảo luận, tức là cần thấu hiểu khách hàng. Trong nghiên cứu này, trước tiên chúng tôi đã thu thập tập ngữ liệu với 26,482 ý kiến nhận xét và bình luận của khách hàng bằng tiếng Anh từ một số trang web thương mại điện tử trong lĩnh vực khách sạn. Sau khi tiền xử lý dữ liệu thu thập được, mô hình được đánh giá thông qua các phép đo Perplexity và Coherence Score để chọn số lượng chủ đề (K) tốt nhất làm tham số đầu vào cho mô hình. Cuối cùng, thực nghiệm trên tập ngữ liệu theo mô hình chủ đề Latent Dirichlet Allocation (LDA) với hệ số K để khám phá chủ đề tiềm ẩn. Kết quả mô hình đã tìm ra các chủ đề ẩn với tập từ khóa tương ứng, đây cũng chính là những thông tin phản ánh những vấn đề khách hàng trong lĩnh vực khách sạn đang quan tâm. Ứng dụng các kết quả thực nghiệm từ mô hình sẽ hỗ trợ cho việc ra quyết định để cải thiện sản phẩm và dịch vụ trong kinh doanh cũng như trong quản lý và phát triển của các doanh nghiệp trong lĩnh vực dịch vụ khách sạn.

Từ khoá: lĩnh vực khách sạn, phân tích dữ liệu, ý kiến khách hàng trực tuyến, mô hình chủ đề

GIỚI THIỆU

Kinh doanh khách sạn là một trong những ngành dịch vụ đặc thù thu được nhiều lợi nhuận của khách hàng, nhưng cũng chịu không ít áp lực cạnh tranh, ý kiến từ khách hàng. Chỉ cần có một số bình luận không hay về khách sạn sẽ làm ảnh hưởng không nhỏ đến hoạt động kinh doanh cũng như phát triển cho cả thời gian dài sau này của doanh nghiệp. Chính vì vậy các doanh nghiệp trong lĩnh vực này cần có phương án để tiếp nhận những phản hồi sau mỗi lần cư trú của khách hàng qua các kênh thông tin bán phòng trực tuyến hay khảo sát trực tiếp ngay chính khách sạn của mình. Cụ thể là các khách sạn có thể lựa chọn những tính năng hay hình thức lấy thông tin phản hồi từ khách hàng khác nhau như: lấy ý kiến trực tiếp, lấy thông tin từ các trang bán hàng trực tuyến, hay lựa chọn một đơn vị thiết kế trang web khách sạn và du lịch có chức năng đánh giá từ khách hàng [1,2].

Hàng ngày có nhiều người dùng mua sản phẩm, đặt vé du lịch, mua hàng hóa và dịch vụ qua web. Người dùng cũng chia sẻ quan điểm của họ về sản phẩm, khách sạn, tin tức và chủ đề trên web dưới dạng đánh giá, blog, nhận xét. Nhiều người dùng đọc thông tin đánh giá được cung cấp trên web để đưa ra quyết định như mua sản phẩm, xem phim, đi ăn nhà hàng. Bài đánh giá chứa ý kiến của người dùng về sản phẩm, sự kiện hoặc chủ đề. Rất khó để người dùng web đọc và hiểu nội dung từ một số lượng lớn các bài đánh giá. Thông tin quan trọng và hữu ích có thể được trích xuất từ các bài đánh giá thông qua quá trình khai thác và tóm tắt ý kiến [3,4]. Điều này đòi hỏi phải có một phương pháp để tổng hợp và trích xuất thông tin từ lượng dữ liệu văn bản này thành các đặc điểm sâu sắc, chẳng hạn như những chủ đề của các bình luận hoặc ý kiến, hoặc những đánh giá trực tuyến về sản phẩm, dịch vụ khách hàng đang nói đến, tức là những "chủ đề" mà họ đang quan tâm.

Phương pháp phân tích ý kiến khách hàng tiếp cận theo mô hình phân tích dữ liệu văn bản và xử lý ngôn ngữ tự nhiên [5] được nhiều công trình nghiên cứu quan tâm. Đặc biệt là các vấn đề về phân tích dữ liệu phi cấu trúc, rút trích thông tin, tóm tắt thông tin. Trong đó, thời gian qua mô hình chủ đề [6] cũng được nhiều tác giả nghiên cứu và thực nghiệm trên tập dữ liệu là các bình luận của khách hàng để lại trên các kênh tiếp nhận phản hồi trực tuyến. Các nghiên cứu này liên quan đến các lĩnh vực như y sinh, giáo dục, nhà ở, mạng xã hội và bán hàng trực tuyến [7–11]. Nhìn chung, nội dung trao đổi của người dùng rất đa dạng phong phú; do đó, đối với các nhà phân tích khi đã khám phá ra các chủ đề nghĩa là khám phá được

[1]*Trường Đại học Kinh tế TP. Hồ Chí Minh, Việt Nam*

[2]*Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam*

Liên hệ

Hồ Trung Thành, Trường Đại học Kinh tế - Luật, ĐHQG-HCM, Việt Nam

Email: thanhht@uel.edu.vn

Lịch sử

- Ngày nhận: 03/09/2020
- Ngày chấp nhận: 26/10/2020
- Ngày đăng: 09/11/2020

DOI :10.32508/stdjelm.v4i4.692

Trích dẫn bài báo này: Hồ N V, Thành H T. **Phân tích ý kiến khách hàng trực tuyến trong lĩnh vực khách sạn tiếp cận theo mô hình chủ đề**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 4(4):1081-1092.

các thông tin quan trọng, cũng như nắm bắt được thói quen, hành vi của người dùng. Tuy nhiên, đối với tính chất của mạng trực tuyến thì chủ đề của nội dung thông điệp trao đổi chưa được tạo trước hay nói cách khác chủ đề được trao đổi trên diễn đàn mạng là tiềm ẩn [10]. Chính vì vậy, việc khám phá chủ đề và hiểu được nội dung thông điệp trao đổi của khách hàng là một thách thức lớn và là bài toán khó [5,9].

Với nghiên cứu trong bài báo này, kết quả của mô hình thực nghiệm trên tập dữ liệu được thu thập, xử lý và tìm ra các chủ đề ẩn mà khách hàng đã trao đổi về các dịch vụ khách sạn, từ đó giúp người quản trị nắm bắt được những vấn đề mà khách hàng quan tâm. Và với những những vấn đề đã tìm ra, với một chiến dịch quảng cáo thông qua thư điện tử chúng ta có thể giữ chân khách hàng, thậm chí với những chiến lược tiếp thị phù hợp chúng ta hoàn toàn có thể nâng cao sự hài lòng của khách hàng hiện có, hay là gia tăng tỉ lệ chuyển đổi thành khách hàng khi áp dụng một chính sách kinh doanh phù hợp với sản phẩm, dịch vụ mục tiêu với đúng khách hàng.

Phần ***Các nghiên cứu liên quan*** gồm là những nghiên cứu liên quan, khảo sát các nghiên cứu về phân tích dữ liệu văn bản, phân tích ý kiến khách hàng trong lĩnh vực khách sạn và mô hình chủ đề LDA. ***Phương pháp nghiên cứu*** được đề cập ở phần tiếp theo. Các vấn đề về thực nghiệm và xây dựng mô hình LDA trên tập dữ liệu được trình bày ở phần ***Đề xuất mô hình nghiên cứu thực nghiệm***. Các chủ đề tìm được và trực quan kết quả sẽ được đề cập và thảo luận trong Phần ***Kết quả nghiên cứu và thảo luận***. Cuối cùng là các ***Kết luận và hướng phát triển***.

CÁC NGHIÊN CỨU LIÊN QUAN

Ngày nay, ngành công nghiệp khách sạn đã trải qua sự tăng trưởng liên tục và phát triển sâu mạnh trên khắp thế giới được thừa nhận bởi các tổ chức quốc tế như Ngân hàng Thế giới và Tổ chức Du lịch Thế giới (WTO) [12]. Chính vì sự tăng trưởng mạnh mẽ của lĩnh vực này và sự phát triển của thương mại điện tử cũng như Internet, khách hàng càng có nhiều lựa chọn hơn khi mua sắm hay sử dụng dịch vụ. Việc hiểu khách hàng là một thách thức lớn đặt ra không chỉ cho các doanh nghiệp kinh doanh dịch vụ khách sạn mà còn cả đối với người quản lý. Chính vì vậy, đã có nhiều nghiên cứu được thực hiện với đa dạng những để các phương pháp và mô hình khác nhau để ứng dụng vào phân tích trải nghiệm khách hàng để nâng cao chất lượng sản phẩm và dịch vụ. Trong đó, lĩnh vực nghiên cứu phân tích và khai thác ý kiến từ đánh giá khách sạn của khách hàng dựa trên các kỹ thuật xử lý ngôn ngữ tự nhiên và học máy [3,12–14]. Trong nghiên cứu của Raut & Londhe [3], tác giả đã trình bày phương pháp học máy và dựa trên SentiWordNet để khai thác ý kiến từ các đánh giá khách sạn và phương pháp dựa trên mức độ liên quan của câu để tổng hợp ý kiến về các đánh giá khách sạn. Dựa trên kết quả nghiên cứu này, thông tin đánh giá khách sạn được phân loại và tóm tắt giúp người dùng web dễ dàng hiểu nội dung đánh giá trong thời gian ngắn. Trong một nghiên cứu khác [4], các tác giả cũng đã phân tích ý kiến phản hồi của khách hàng trong lĩnh vực du lịch bằng cách để xuất một kỹ thuật tóm tắt đa văn bản mới để xác định các câu thông tin nhất trong các bài đánh giá về khách sạn. Trong nghiên cứu của Hu *et. al* [4] cũng đã xem xét các yếu tố về sự giống nhau về nội dung và tình cảm và được sử dụng để xác định sự giống nhau của hai câu bình luận. Thuật toán phân cụm k-medoids được sử dụng để phân chia các câu thành k nhóm. Medoids từ các nhóm này sau đó được chọn làm kết quả tổng hợp cuối cùng. Để đánh giá hiệu suất của phương pháp để xuất, nhóm tác giả đã thu thập hai bộ đánh giá cho hai khách sạn được đăng trên TripAdvisor.com. Tổng số 20 đối tượng đã được mời để xem xét các kết quả tóm tắt văn bản từ cách tiếp cận để xuất và hai cách tiếp cận thông thường cho hai khách sạn. Kết quả chỉ ra rằng cách tiếp cận được để xuất vượt trội hơn hai cách còn lại và hầu hết các đối tượng tin rằng cách tiếp cận được để xuất có thể cung cấp thông tin khách sạn toàn diện hơn.

Trong nghiên cứu của Berezina *et. al* [15], tác giả xem xét những cơ sở nền tảng của khách hàng hài lòng và không hài lòng thông qua phương pháp phân tích văn bản. Đánh giá trực tuyến của 2,510 khách khách sạn đã được thu thập từ TripAdvisor.com cho Sarasota, Florida. Kết quả nghiên cứu cho thấy một số "chủ đề" phổ biến được sử dụng trong cả đánh giá tích cực và tiêu cực, bao gồm địa điểm kinh doanh (ví dụ: khách sạn, và câu lạc bộ), phòng, nội thất, thành viên và thể thao. Kết quả nghiên cứu cũng chỉ ra rằng những khách hàng hài lòng sẵn sàng giới thiệu khách sạn cho người khác để cập đến những khía cạnh vô hình trong việc lưu trú tại khách sạn của họ, chẳng hạn như nhân viên, thường xuyên hơn những khách hàng không hài lòng. Mặt khác, những khách hàng không hài lòng để cập thường xuyên hơn đến các khía cạnh hữu hình của khách sạn, chẳng hạn như nội thất và tài chính (chi phí, giá cả). Nghiên cứu đưa ra các hàm ý lý thuyết và quản lý rõ ràng liên quan đến việc hiểu khách hàng hài lòng và không hài lòng thông qua việc sử dụng khai thác văn bản và xếp hạng khách sạn thông qua các trang web đánh giá, phương tiện truyền thông xã hội, blog và các nền tảng trực tuyến khác.

PHƯƠNG PHÁP NGHIÊN CỨU

Trong các nghiên cứu về phân tích ý kiến khách hàng[9,11], các tác giả cũng thực hiện nghiên cứu thực nghiệm trên tập dữ liệu phi cấu trúc là các bình luận của khách hàng. Dữ liệu này được thu thập chủ yếu thông qua các kênh trực tuyến và các công cụ thu thập ý kiến, các bảng khảo sát đánh giá của doanh nghiệp. Một số nghiên cứu cũng đã quan tâm đến lĩnh vực nhà hàng khách sạn[13,14]. Cụ thể hơn, các phương pháp phân tích dữ liệu văn bản, mô hình chủ đề là một trong những cách tiếp cận hiệu quả trong việc tìm ra các chủ đề tiềm ẩn từ tập khổng lồ là các phản hồi trực tuyến của khách hàng [12]. So với tập các ý kiến ban đầu, kết quả thực nghiệm của các nghiên cứu này là tập chủ đề và tập từ khóa được thể hiện ngắn gọn và rõ ràng hơn.

Phương pháp khai phá văn bản

Khai phá văn bản, còn được gọi là khai phá dữ liệu văn bản, tương tự như phân tích văn bản, là quá trình lấy thông tin chất lượng cao từ văn bản[16,17]. Khai thác văn bản là một phần quan trọng của quá trình khai thác dữ liệu và khám phá tri thức, liên quan đến việc phát hiện ra thông tin mới, trước đây chưa được biết đến, bằng cách tự động trích xuất thông tin từ các nguồn tài liệu viết khác nhau. Các nguồn tài liệu viết có thể bao gồm trang web, sách, email, các đánh giá bình luận và bài báo. Thông tin chất lượng cao thường thu được nhờ vào sử dụng kỹ thuật là các thuật toán khai thác dữ liệu như thống kê và học máy. Có thể phân biệt ba quan điểm khác nhau của khai thác văn bản: khai thác thông tin, khai thác dữ liệu và khám phá tri thức (KDD – Knowledge Discovery in Databases)[16]. Các tác vụ khai thác văn bản điển hình bao gồm phân loại văn bản, phân cụm văn bản, trích xuất khái niệm – thực thể, tìm ra các đơn vị phân loại chi tiết, phân tích tình cảm, tóm tắt tài liệu và mô hình hóa quan hệ thực thể. Về cơ bản, mục tiêu bao trùm là biến văn bản thành dữ liệu để phân tích, thông qua ứng dụng xử lý ngôn ngữ tự nhiên (NLP – Natural Language Processing), các loại thuật toán và phương pháp phân tích. Một giai đoạn quan trọng của quá trình này là giải thích thông tin thu thập được[18].

Ý kiến khách hàng là những phản hồi, khen chê, góp ý mà khách hàng đưa ra sau khi sử dụng sản phẩm hay thương hiệu của công ty. Phân tích khai thác ý kiến khách hàng là nghiên cứu phân tích ý kiến, tình cảm, đánh giá, thái độ và cảm xúc của mọi người từ ngôn ngữ viết. Hiện nay, với sự phát triển của công nghệ và nền tảng di động trực tuyến, người dùng có thể dễ dàng đưa ra nhận xét của mình về chất lượng dịch vụ phòng, dịch vụ khách hàng. Khách hàng có thể đính kèm các hình ảnh thực tế về sản phẩm và dịch vụ nhận được vào các bình luận để minh chứng cho nhận xét của mình trở nên đáng tin cậy và thuyết phục hơn. Có thể nhận thấy, với sự phát triển nhanh chóng như vậy trong thời đại kỹ thuật số phát triển, chúng ta hiện có một khối lượng dữ liệu lớn được ghi lại dưới dạng "kỹ thuật số" để phân tích. Đây cũng chính là một trong những động lực dẫn đến nghiên cứu trong bày báo này được thực hiện.

Mô hình chủ đề LDA

Mô hình chủ đề LDA là một mô hình xác suất được áp dụng để mô hình hóa nhằm khám phá ra các chủ đề ẩn từ kho ngữ liệu[6]. Ngược lại với quá trình tạo thông điệp, mô hình LDA thực hiện trên sự đồng hiện của tập từ trong ngữ liệu để gom cụm các từ. Trong học máy và xử lý ngôn ngữ tự nhiên, mô hình chủ đề là một mô hình thống kê để khám phá các cấu trúc ngữ nghĩa ẩn dựa trên các biến ẩn của mô hình, các "chủ đề" trừu tượng xảy ra trong một bộ tài liệu văn bản. Hình 1 biểu diễn minh họa cho tiến trình sinh xác suất giữa văn bản, từ, và chủ đề trong mô hình. Kết quả của LDA bao gồm phân phối xác suất theo văn bản và phân phối xác suất theo từ.

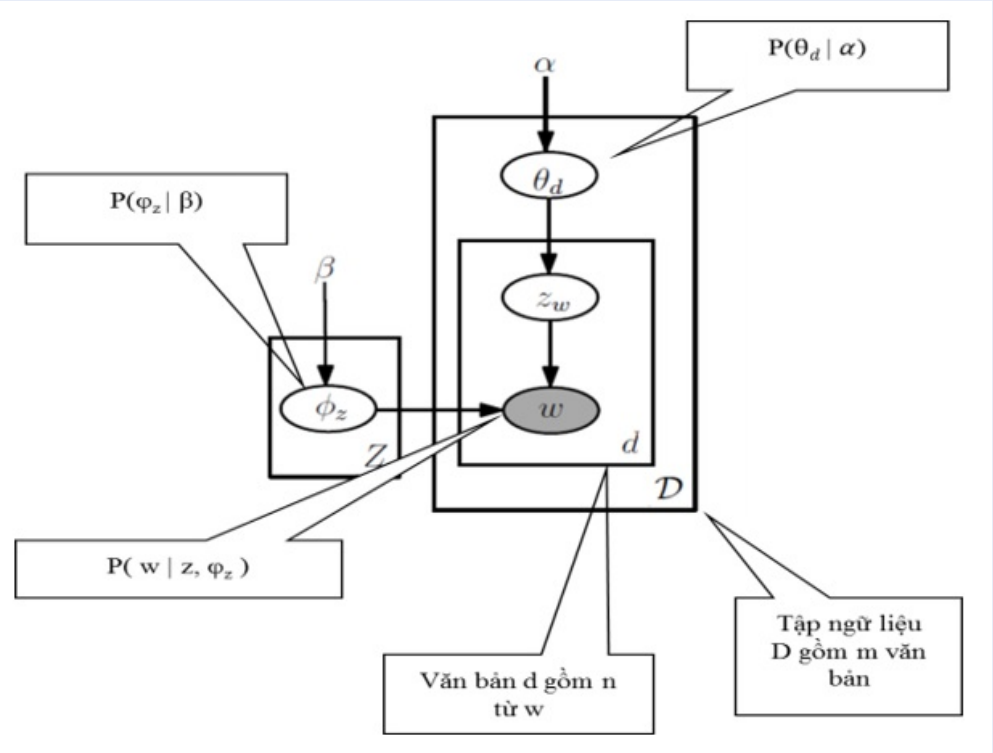
Bảng 1 là mô tả các định nghĩa và ký hiệu sử dụng trong mô hình LDA. Ở đây, có hai quá trình lặp lại liên tục trong LDA là quá trình lựa chọn chủ đề và quá trình lựa chọn từ. Các tham số được khởi tạo tiến trình ban đầu là α và β. Từ đó tính toán được phân phối hỗn hợp của chủ đề θ và phân phối của từ theo chủ đề w.

Kỹ thuật lấy mẫu Gibbs cho mô hình chủ đề LDA

Các biến ẩn trong mô hình LDA[6] trên bao gồm chủ đề z, phân bố từ trong chủ đề $\varnothing$, phân bố chủ đề trong thông điệp θ. Phân bố hậu nghiệm của các biến này được phân tích bằng cách sử dụng lý thuyết Bayes. Xét theo từng từ w, ta tính tổng xác suất của mô hình dựa trên từng từ w và từ đó suy ra tổng xác suất của mô hình trên cả kho ngữ liệu D. Trong mô hình LDA, các đại lượng biến ẩn này được tính theo công thức sau:

$$P(\theta,\varnothing,z|w;\alpha,\beta)=\frac{P(\theta,\varnothing,z,w|\alpha,\beta)}{P(w|\alpha,\beta)}=\frac{P(\theta,\varnothing,z,w|\alpha,\beta)}{\int_\theta\int_\varnothing\sum_{i=1}^{K}P(w,z,\theta,\varnothing|\alpha,\beta)d\varnothing d\theta} \quad (1)$$

Tuy nhiên, các yếu tố chuẩn hóa $P(w|\alpha,\beta)$ (hay phân phối biên) không thể tính một cách chính xác[6,19] vì $P(w|\alpha,\beta)$ không đổi cho bất kỳ chủ đề z nào hay nói cách khác không thể tính biên qua các biến ẩn. Việc áp dụng phương pháp lấy mẫu, phân bố hậu nghiệm



Hình 1: Mô hình LDA [6]

Bảng 1: Các định nghĩa sử dụng trong mô hình LDA

Thứ tự	Ký hiệu	Diễn giải
1	W	Từ, có chỉ số trong từ điển có giá trị từ 1,2, …, V
2	M	Số lượng văn bản
3	D	Tập hợp M văn bản – tập ngữ liệu
4		Số lượng từ trong mỗi văn bản M
5	α	Tham số của phân phối tiên nghiệm Dirichlet của chủ đề đối với mỗi văn bản
6	β	Tham số của phân phối tiên nghiệm Dirichlet của chủ đề đối với mỗi từ
7	θ_i	Phân phối của chủ đề đối với văn bản thứ i
8	φ_k	Phân phối của từ đối với chủ đề thứ k
9	z_{ij}	Chủ đề của từ thứ j đối với văn bản thứ i, là số nguyên trong khoảng [1, K]
10	w_{ij}	Chỉ số trong bộ từ của từ thứ i thuộc văn bản thứ j, là số nguyên trong khoảng [1, V]
11	K	Số lượng chủ đề

cho (1) được tính xấp xỉ thông qua các mẫu của phân bố xác suất liên hợp được trình bày trong (2).

$$P(\theta, \varnothing, z|w; \alpha, \beta) = \frac{P(\theta, \varnothing, z, w|\alpha, \beta)}{P(w|\alpha, \beta)} \propto P(\theta, \varnothing, z, w|\alpha, \beta) \quad (2)$$

Nhìn chung, việc thực hiện lấy mẫu Gibb cho tất cả các biến trong mô hình LDA là khả thi [20]. Tuy nhiên, việc đó lại không hiệu quả bởi vì việc lấy mẫu cho tham số đa thức θ và $\varnothing$ được tính từ các biến chủ đề z mà z lại là biến ẩn. Nói cách khác, việc thực hiện lấy mẫu Gib s nên được thực hiện bằng cách kết hợp giữa phân bố Dirichlet và phân bố xác suất nhiều chiều để tính tích phân theo các tham số đa thức θ và $\varnothing$ trong công thức (2) và áp dụng giải thuật Collapsed Gibbs sampling [20] được dùng kết hợp với mô hình LDA [6] để tính xác suất của một chủ đề z đang được gán vào từ w_i dựa theo tất cả các phép gán của chủ đề z khác vào các từ w khác, tức là tính:

$P(z_i|z_{-i}, \alpha, \beta, w)$.

Dưới đây là giải thuật lấy mẫu Gibbs cho mô hình LDA [20] và áp dụng phương pháp trong nghiên cứu của Roy Daniel and Sontag David [19] để tính toán độ phức tạp của Bảng 2.

ĐỀ XUẤT MÔ HÌNH NGHIÊN CỨU THỰC NGHIỆM

Mô hình nghiên cứu tổng quan

Khai phá ý kiến có nghĩa là tìm và phân loại các phần có ý kiến của văn bản. Những phần chủ quan này cần được xác định bằng các phương pháp khai phá văn bản và được tách biệt khỏi các phần văn bản khách quan. Khai phá ý kiến có thể được coi là một quá trình với ba mức phân loại chính: mức tài liệu (document level), mức câu văn (sentence level) và mức khía cạnh (aspect level) [21]. Để tìm ra các chủ đề phổ biến mà khách hàng thương quan tâm, trong nghiên cứu này trước hết chúng tôi tiến hành thu thập các ý kiến đánh giá hay nhận xét về một vấn đề nào đó, sau đó trích lọc các ý kiến viết bằng tiếng Anh. Tập dữ liệu này sẽ được tiền xử lý thông qua các gói công cụ hỗ trợ từ thư viện của Python. Sau khi đánh giá mô hình tìm ra số chủ đề tối ưu làm làm số đầu vào cho mô hình LDA, chúng tôi tiến hành chạy mô hình thực nghiệm. Các chủ đề ẩn được tìm thấy và biểu diễn trực quan hóa. Hình 2 trình bày mô hình nghiên cứu thực nghiệm được đề xuất từ giai đoạn thu thập xử lý, xây dựng mô hình LDA, phân tích ý kiến khách hàng và trực quan hóa kết quả.

Thu thập dữ liệu

Dữ liệu sử dụng trong bài viết này được thu thập từ các trang web trong lĩnh vực khách sạn, cụ thể là trang web https://www.agoda.com/. Để thu thập dữ liệu, nhóm nghiên cứu lập trình ứng dụng, sử dụng thư viện Selenium của Python để truy cập vào API của website và thu thập các nhận xét và bài viết của khách hàng trên các trang đánh giá lưu thành các tập tin với định dạng JSON. Sau đó, chuỗi dữ liệu JSON được chuyển sang định dạng dữ liệu CSV và thực hiện phân tích rút trích chủ đề trên tập dữ liệu thu thập được. Một số thuộc tính được rút trích để phân tích bao gồm *hotel_id*, *review_comments*, *language_comments, review_date*. Tổng số 26,482 ý kiến nhận xét của khách hàng đã được thu thập, sau đó chúng được sử dụng làm đầu vào để phân tích ý kiến của khách hàng.

Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một trong những bước quan trọng nhất trong khai thác dữ liệu, đặc biệt là trong khai thác dữ liệu văn bản vì có rất nhiều sự khác biệt về nội dung văn bản trên các kênh truyền thông điện tử như trên Internet.

Những bình luận và ý kiến của khách hàng sử dụng sản phẩm và dịch vụ khách sạn thông qua nội dung văn bản trên các trang thương mại điện từ thường chứa đựng hoặc lặp lại một số kí tự đặc biệt hay từ viết tắt để nhấn mạnh các thông điệp của họ. Cách diễn đạt này có thể gây nhập nhằng và khó khăn cho các mô hình phân tích ý kiến của khách hàng, để tránh vấn đề này xảy ra trong quá trình xử lý, các kí tự hay từ viết tắt đặc biệt trong các bình luận sẽ được loại bỏ hoặc được ánh xạ sang từ rõ nghĩa hơn. Các dấu chấm câu không có ý nghĩa trong bộ dữ liệu cũng sẽ bị xóa. Các ký tự viết hoa sẽ được chuyển đổi thành chữ thường, loại bỏ số và khoảng trống, và các từ dừng (stop word) cũng đảm bảo được loại bỏ. Với xu hướng phát triển của thiết bị di động thông minh dẫn đến nhiều ứng dụng di động cũng được phát triển theo. Điều này dẫn đến nhiều khách hàng truy cập các dịch vụ mạng xã hội, trang thương mại điện tử qua điện thoại di động và có xu hướng bỏ qua các quy tắc ngữ pháp và chính tả, sử dụng các chữ viết tắt, biểu tượng cảm xúc và các câu ngắn gọn hơn. Chính vì vậy, giai đoạn thu thập và tiền xử lữ liệu là rất quan trọng và là một trong những yếu tố xử lý dữ liệu nhiễu và tham gia vào việc quyết định tính chính xác của mô hình. Hình 3 dưới đây là qui trình tiền xử lý dữ liệu trước khi đưa vào xây dựng mô hình LDA.

Xây dựng mô hình LDA

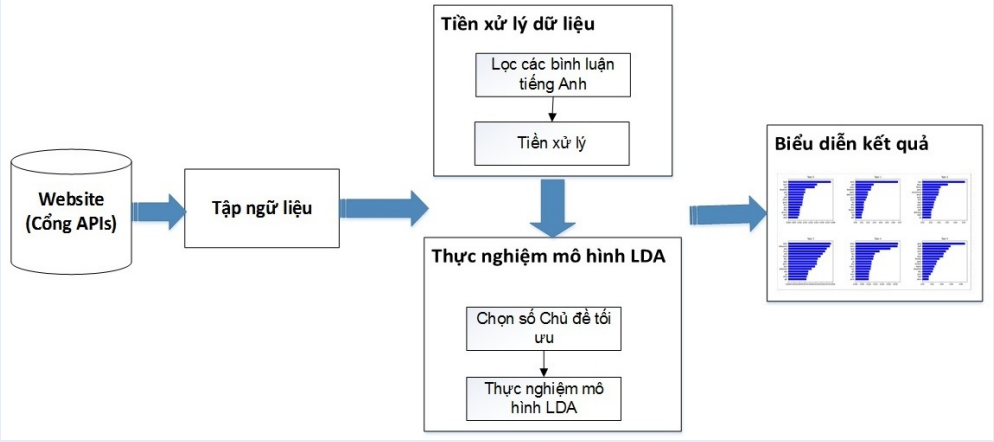
Trong quá trình xây dựng mô hình LDA, có 3 bước quan trọng để thực hiện. Trong đó gồm:

Bảng 2: Giải thuật lấy mẫu Gibbs cho mô hình LDA [20] và độ phức tạp

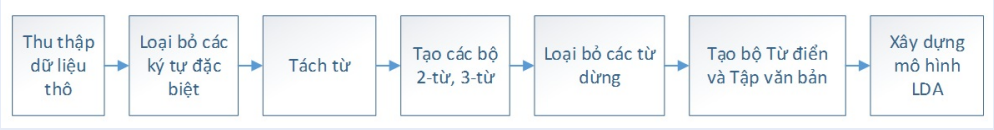
	Giải thuật lấy mẫu Gibbs cho mô hình LDA [20] và độ phức tạp	Giải thích
1	Đầu vào: kho ngữ liệu thông điệp *D*, số lượng chủ đề k cần khám phá, tham số Dirichlet α, β	
2	Đầu ra: các phép gán chủ đề và các biến đếm $n_{d,k}$, $n_{k,w}$, n_k. Bao gồm ∅ phân bố tập từ w trong chủ đề z, θ phân bố chủ đề z trong thông điệp d	
3	Bắt đầu	
4	Khởi tạo biến ngẫu nhiên z và lặp các biến đếm	
5	foreach bước lặp do // lặp từng thông điệp *d* thuộc tập thông điệp *D*	*M* là số thông điệp trong kho ngữ liệu *D*.
6	for i = 0 $\rightarrow N$ – 1 do // lặp từng từ trong mỗi thông điệp d	
7	từ ← w[i]	
8	chủ đề ← z[i]	
9	$n_{d,\ ch\ đ-} = 1$; $nt_{,\ ch\ đ-} = 1$; $n_{ch\ đ-} = 1$	
10	for k =0 → K -1 do // lặp theo số lượng chủ đề cần rút trích *// tính xác suất của chủ đề z đang được gán vào từ w dựa vào tất cả các phép gán của các chủ đề z khác vào các từ w khác*	*N* là số từ của mỗi thông điệp *d*.
11	$P(z = k\|\bullet) = (n_{d,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$	*K* là số lượng chủ đề cần khám phá.
12	end	
13	chủ đề ← lấy mẫu từ $p(z\|\bullet)$	
14	z[i] ← chủ đề	
15	$n_{d,\ ch\ đ+} = 1$; $n_{t,\ ch\ đ+} = 1$; $n_{ch\ đ+} = 1$	*I* là số lần lặp lấy mẫu Gibbs cho LDA.
16	end	
17	end	
18	return z, $n_{d,k}$, $n_{k,w}$, n_k	
19	Kết thúc	

Độ phức tạp giải thuật được tính toán dựa trên bốn vòng lặp tại:
- Dòng 5: lặp theo mỗi thông điệp d trong kho ngữ liệu *D*
- Dòng 6: lặp theo *N* từ trong từng thông điệp d thuộc tập ngữ liệu *D*
- Dòng 10: lặp theo số lượng chủ đề *K*
- Dòng 13: lấy mẫu Gibbs và lặp theo chỉ số *I*.

Tổng chi phí thời gian thực hiện của giải thuật lấy Gibbs cho LDA là: *D*N*K*I*
Từ đó suy ra độ phức tạp của giải thuật là: *O(D*N*K*I)*



Hình 2: Mô hình nghiên cứu thực nghiệm tổng quan



Hình 3: Qui trình Tiền xử lý dữ liệu

1) Tạo n-gram: Mô hình LDA sẽ sử dụng đầu vào là ma trận đồng xuất hiện của các từ. Để tính toán được tần suất đồng xuất hiện trên những ma trận này chúng ta sẽ tạo ra các bộ 2-từ (bigram) và 3-từ (trigram) là cụm các từ liên tiếp nhau. Hàm *class_phrases()* của Gensim được sử dụng để xây dựng các bộ 2-từ và 3-từ. Tham số *min_count* chính là tần suất nhỏ nhất để một từ được lựa chọn đưa vào các gram và ngưỡng cho phép được thiết lập. Tiếp theo các từ dừng (stop-words) sẽ được loại bỏ và chỉ lọc ra các từ vựng là các từ có thuộc từ loại là danh từ, tính từ, trạng trừ, và động từ. Bộ từ dừng trong tiếng Anh đã được tích hợp sẵn trong gói *nltk*[a];

2) Tạo ra từ điển và bộ văn bản: Từ điển (dictionary) và bộ văn bản (corpus) là hai yếu tố đầu vào chính cho mô hình LDA. Gói Gensim được sử dụng để tạo chúng. Sau khi xử lý ta đã thu được một bộ văn bản là tập các cặp (chỉ số, tần suất) mã hóa các văn bản về chỉ số được qui định trong từ điển kèm theo tần suất xuất hiện của chúng trong văn bản;

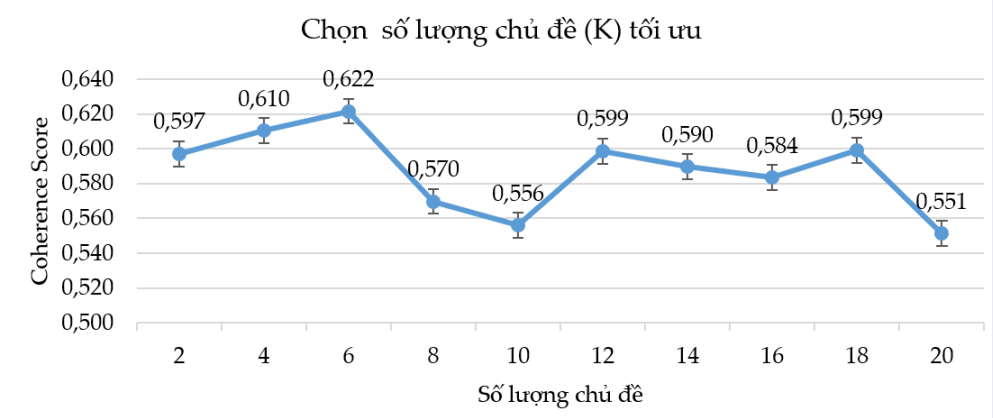
3) Lựa chọn số chủ đề K: Mô hình LDA được huấn luyện với mục đích các đoạn văn bản được biểu diễn bằng một số các chủ đề và các chủ đề đó lại được biểu diễn bằng một tập các từ, với trọng số ứng với từng từ giảm dần. Tham số chính được qui định trong mô hình LDA chính là số lượng chủ đề *K*, số lượng văn bản được đưa vào mỗi lần huấn luyện (chunksize), số lượt huấn luyện (passes). Trong nghiên cứu này, chúng tôi đã thực nghiệm và chọn ra số chủ đề là 6 (với các chỉ số tương ứng Perplexity: - 6.839 và Coherence Score: 0.622) để làm tham số đầu vào cho mô hình. Hình 4 là biểu đồ thể hiện sự tương quan giữa chỉ số Coherence Score (CS) và số lượng chủ đề, dựa vào chỉ số CS cao nhất để chọn ra số chủ đề K tối ưu.

[a]https://www.nltk.org/ (truy cập ngày 01/9/2020)

KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

Tập chủ đề

Kết quả thực nghiệm mô hình LDA với chỉ số K, tham số được khởi tạo tiến trình ban đầu là α và β đã tìm ra các chủ đề cùng với xác suất sinh tương ứng của từ trong chủ đề đó (ma trận chủ đề - từ với xác suất tương ứng). Các chủ đề chiếm ưu thế trong tập văn bản được đề xuất, tức là những chủ đề có tỉ lệ xác suất cao nhất. Bảng 3 trình bày trên đã thể hiện tập các từ của từng chủ đề 0, 2, 4 và 5. Ở đây, quan sát chúng ta có thể thấy đối với chủ đề 0 và chủ đề 4, từ "hotel" có xác suất cao nhất là 0.0381 và 0.0320 theo thứ tự tương ứng, tương tự với chủ đề 5, từ "good" có xác suất cao nhất với giá trị là 0.0883. Tuy nhiên, trong chủ đề 5, từ "helpful" lại ít được khách hàng sử dụng dịch vụ để



Hình 4: Tương quan giữa Coherence Score và Số lượng chủ đề (K)

Bảng 3: Các chủ đề 0, 2, 4 và 5 cùng với mười từ có xác suất cao nhất

Chủ đề 0		Chủ đề 2		Chủ đề 4		Chủ đề 5	
Từ	Xác suất	Từ	Xác suất	Từ	Xác suất	Từ	Xác suất
hotel	0.0381	stay	0.0517	hotel	0.0320	good	0.0883
room	0.0257	place	0.0308	room	0.0316	staff	0.0586
bad	0.0236	family	0.0220	book	0.0264	hotel	0.0543
breakfast	0.0151	great	0.0202	staff	0.0182	room	0.0487
pool	0.0150	recommend	0.0199	bus	0.0146	nice	0.0441
old	0.0138	would	0.0192	day	0.0137	clean	0.0430
staff	0.0135	really	0.0168	give	0.0134	friendly	0.0355
check	0.0130	time	0.0159	go	0.0121	great	0.0301
guest	0.0127	back	0.0144	pay	0.0120	location	0.0286
time	0.0124	make	0.0137	check	0.0117	helpful	0.0252

cập đến chỉ với xác suất "0.0252". Hoặc trong chủ đề 0, từ "old" được khách hàng dùng để đánh giá dịch vụ của khách sạn với xác suất 0.0138 ở mức trung bình trong 10 từ nổi bật thuộc chủ đề.

Suy luận nhãn chủ đề

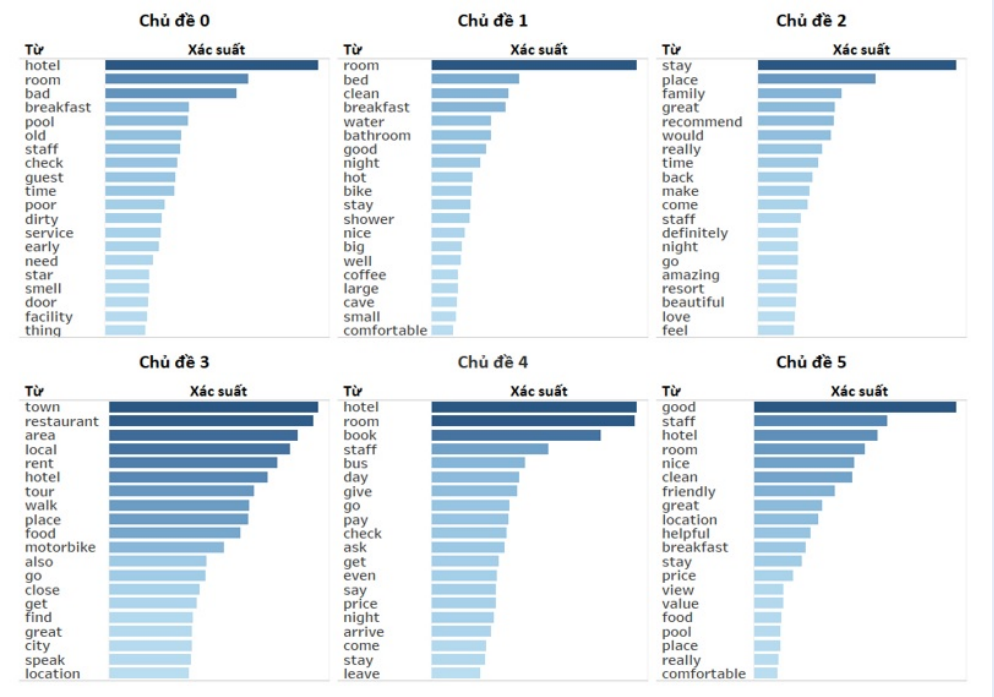
Hình 5 trình bày các chủ đề chiếm ưu thế trong tập ngữ liệu và là tập hợp những từ có tỉ lệ xác suất cao nhất. Các chủ đề tìm được và bộ 10 từ với tần suất cao của mỗi chủ đề. Từ tập từ khóa này, với tập từ khóa này, chúng ta có thể suy luận nhãn của chủ đề 0 là "hotel_services". Cũng như vậy, nhãn của chủ đề 1 là "room_types".

Biểu diễn trực quan

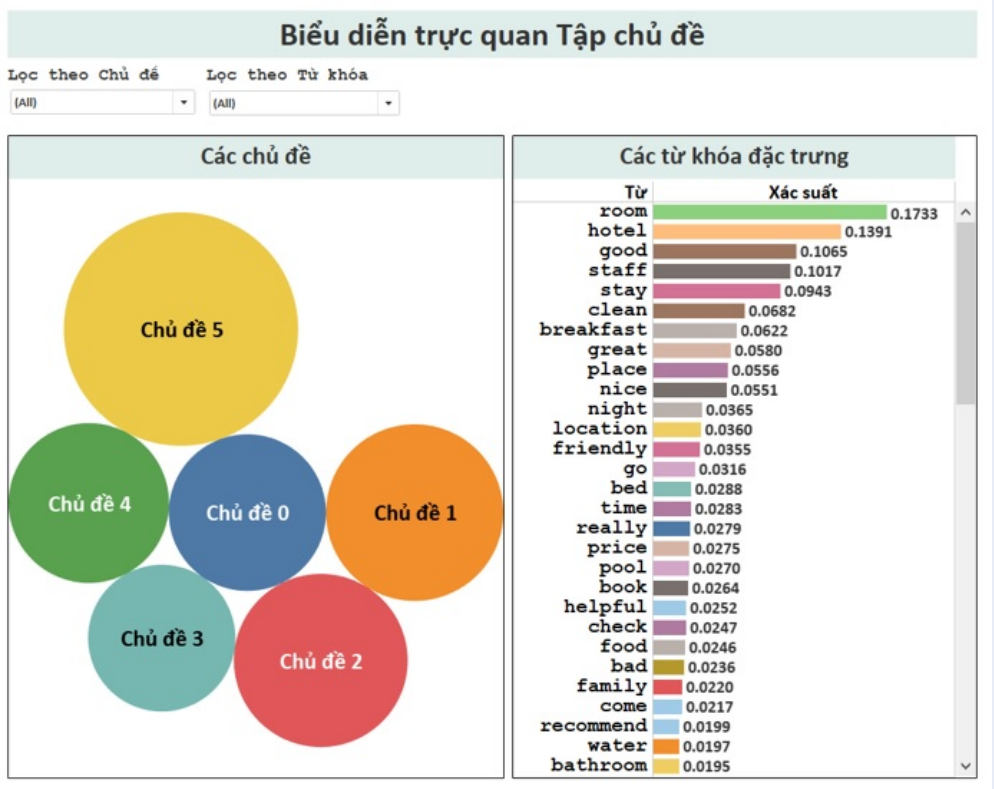
Hình 6 là kết quả của mô hình thực nghiệm được biểu diễn trực quan hóa. Có thể thấy, biểu đồ này có thể tương tác trực tiếp để lựa chọn những yếu tố cần phân tích. Một chủ đề trong tập kết quả được biểu diễn bởi một hình tròn. Hình tròn có bán kính càng lớn, chủ đề đó càng ưu thế (được quan tâm nhiều). Chúng ta có thể di chuyển con trỏ qua một trong các hình tròn khác, các từ đặc trưng và thanh biểu diễn xác suất ở phía bên phải sẽ cập nhật. Những từ này là các từ khóa nổi bật tạo thành chủ đề được chọn. Các bộ lọc theo chủ đề và lọc theo từ khóa giúp người dùng báo cáo thuận tiện hơn trong việc phân tích kết quả của mô hình thực nghiệm.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Tóm lại, kinh doanh khách sạn là một trong những ngành dịch vụ đặc thù thu được nhiều lợi nhuận từ khách hàng, nhưng cũng chịu không ít áp lực cạnh tranh từ đối thủ, và nhiều ý kiến phản hồi từ khách



Hình 5: Tập chủ đề được phân tích và các từ khóa đại diện với xác suất cao



Hình 6: Biểu diễn trực quan các chủ đề và các từ khóa với xác suất đồng hiện

hàng. Chính vì vậy, mô hình chủ đề được đề xuất trong nghiên cứu này đã phần nào giải quyết được bài toán thu thập và phân tích ý kiến khách hàng. Trong mô hình thực nghiệm, chúng tôi sử dụng dữ liệu được thu thập từ trang thương mại điện tử Agoda trong khoảng thời gian từ năm 2012 đến năm 2018, dữ liệu này có thể chưa phản ánh toàn diện hiện trạng ý kiến khách hàng về các thương hiệu khách sạn hiện tại nhưng có thể làm đại diện để thực nghiệm mô hình. Kết quả đã cho thấy được tập chủ đề và các các từ khóa trích xuất được đã phản ánh chính xác những vấn đề mà người dùng trong lĩnh vực khách sạn thường quan tâm. Các biểu diễn trực quan kết quả bằng đồ thị và biểu đồ động giúp nhà quản trị nắm bắt thông tin một cách thuận tiện và kịp thời, cho phép họ nhìn vấn đề với các góc nhìn (chiều phân tích) khác nhau.

Trong thời gian sắp tới, đề tài sẽ được phát triển theo hướng xây dựng và đề xuất mô hình thu thập và phân loại ý kiến khách hàng theo thời gian thực và sau đó kết quả phân loại sẽ được đưa tiếp đến các hệ thống phân tích trực tuyến trong đó mỗi bình luận sẽ được ghi nhận cùng với yếu tố thời gian. Hệ thống phân tích ý kiến khách hàng sẽ có thể thực hiện phân tích những thay đổi tiêu cực, tích cực, các vấn đề khách hàng đang phản hồi theo thời gian, từ đó giúp doanh nghiệp nhanh chóng đưa ra chiến lược thích hợp để kịp thời xử lý khủng hoảng hoặc nhận ra và tăng cường các yếu tố làm nâng cao sự hài lòng của khách hàng.

DANH MỤC CÁC TỪ VIẾT TẮT

API: Application Programming Interface
CS: Coherence Score
CSV: Comma-Separated Values
JSON: JavaScript Object Notation
HTML: Hypertext Markup Language
KDD: Knowledge Discovery in Databases
LDA: Latent Dirichlet Allocation
Ngữ liệu (text corpus): một tập dữ liệu tập hợp các văn bản, ngôn ngữ đã được số hoá, một tài nguyên quan trọng trong xử lý ngôn ngữ tự nhiên.
NLP: Natural Language Processing
WTO: World Tourism Organization

XUNG ĐỘT LỢI ÍCH

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

ĐÓNG GÓP CỦA TÁC GIẢ

Toàn bộ nội dung bài viết chỉ do nhóm tác giả thực hiện. Các tác giả có đóng góp như nhau trong quá trình nghiên cứu về ý tưởng, mục tiêu, phương pháp nghiên cứu, đề xuất mô hình, phân tích dữ liệu, đánh giá và thảo luận kết quả.

TÀI LIỆU THAM KHẢO

1. Khoa DL, Ngọc NT. Ảnh hưởng của đánh giá trực tuyến đến quyết định lựa chọn khách sạn của khách du lịch khi đến Huế. Hue University Journal of Science: Economics and Development. 2017;126(5D):41–51. Available from: https://doi.org/10.26459/hueuni-jed.v126i5D.4489.
2. Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? Journal of interactive marketing. 2004;18(1):38–52. Available from: https://doi.org/10.1002/dir.10073.
3. Raut VB, Londhe DD. Opinion mining and summarization of hotel reviews. In2014 International Conference on Computational Intelligence and Communication Networks. IEEE. 2014;p. 556–559. Available from: https://doi.org/10.1109/CICN.2014.126.
4. Hu YH, Chen YL, Chou HL. Opinion mining from online hotel reviews-a text summarization approach. Information Processing & Management. 2017;53(2):436–449. Available from: https://doi.org/10.1016/j.ipm.2016.12.002.
5. Boyd-Graber JL, Hu Y, Mimno D. Applications of topic models. Publishers Incorporated. 2017;Available from: https://doi.org/10.1561/9781680833096.
6. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3:993–1022.
7. Kho SJ, Yalamanchili HB, Raymer ML, Sheth AP. A novel approach for classifying gene expression data using topic modeling. InProceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2017;p. 388–393. Available from: https://doi.org/10.1145/3107411.3107483.
8. Ho T, Do P. An integrated model for discovering, classifying and labeling topics based on topic modeling. Science and Technology Development Journal. 2014;17(2):73–85. Available from: https://doi.org/10.32508/stdj.v17i2.1361.
9. Sutherland I, Kiatkawsin K. Determinants of Guest Experience in Airbnb: A Topic Modeling Approach Using LDA. Sustainability. 2020;12(8):3402. Available from: https://doi.org/10.3390/su12083402.
10. Nguyen M, Ho T, Do P. Social networks analysis based on topic modeling. InThe 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF) . IEEE. 2013;119(122).
11. Moghaddam S, Ester M. ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. InProceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011;p. 665–674. Available from: https://doi.org/10.1145/2009916.2010006.
12. Putri I, Kusumaningrum R. Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia. Journal of Physics: Conference Series. 2017;801:012073. Available from: https://doi.org/10.1088/1742-6596/801/1/012073.
13. Rossetti M, Stella F, Zanker M. Analyzing user reviews in tourism with topic models. Information Technology & Tourism. 2016;16(1):5–21. Available from: https://doi.org/10.1007/s40558-015-0035-y.
14. Shi HX, Li XJ. A sentiment analysis model for hotel reviews based on supervised learning. In 2011 International Conference on Machine Learning and Cybernetics. IEEE. 2011;3:950–954. Available from: https://doi.org/10.1109/ICMLC.2011.6016866.
15. Berezina K, Bilgihan A, Cobanoglu C, Okumus F. Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. Journal of Hospitality Marketing & Management. 2016;25(1):1–24. Available from: https://doi.org/10.1080/19368623.2015.983631.
16. Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. InLdv Forum. 2005;20(1):19–62.

17. Mandl T. Text mining. InEncyclopedia of Information Science and Technology, Third Edition. IGI Global. 2015;p. 1923–1930. Available from: https://doi.org/10.4018/978-1-4666-5888-2.ch185.
18. Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press. 2007;Available from: https://doi.org/10.1017/CBO9780511546914.
19. Daniel R, David S. Complexity of Inference in Latent Dirichlet Allocation, 25th Annual Conference on Neural Information Processing Systems, NIPS 2011 - Granada, Spain;Available from: 2011.
20. Tom Griffiths. Gibbs Sampling in the generative model of Latent Dirichlet Allocation, Gruffydd@psych.stanford.edu. 2004;.
21. Bakshi RK, Kaur N, Kaur R, Kaur G. Opinion mining and sentiment analysis. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) . IEEE. 2016;p. 452–455.

Science & Technology Development Journal – Economics - Law and Management, 4(4):1081-1092

Open Access Full Text Article

Topic modeling for analyzing online reviews in hotel sector

Nguyen Van Ho[1], Ho Trung Thanh[2,*]

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

ABSTRACT

Recently, with the growth of technology and the Internet, customers can easily create their opinions and feedbacks about products and services of hotels on websites or social media. This information is stored in textual form, and is a huge source of data to explore. In order to continue developing to meet customers' needs, businesses need to gain customers' insights that customers discuss and concern. In this study, we firstly collected a corpus of 26,482 customer comments and reviews written in English from some e-commerce websites in the hospitality industry. After preprocessing the collected data, our team conducted experiments on this corpus and chose the best number of topics (K) by Coherence Score measurements as input parameters for the model. Finally, experiment on the corpus according to the Latent Dirichlet Allocation (LDA) model with K coefficient to explore the topic. The model results found hidden topics with the corresponding list of keywords, reflecting the issues that customers are interested in. Applying empirical results from the model will support decision making to improve products and services in business as well as in the management and development of businesses in the hotel sector.

Key words: hotel sector, analyzing data, online reviews, topic modeling

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

[2]*University of Economics and Law, VNU-HCM, Vietnam*

Correspondence

Ho Trung Thanh, University of Economics and Law, VNU-HCM, Vietnam

Email: thanhht@uel.edu.vn

History

- Received: 03/9/2020
- Accepted: 26/10/2020
- Published: 09/11/2020

DOI :10.32508/stdjelm.v4i4.692

Check for updates

VNU-HCM Press

Cite this article : Ho N V, Thanh H T. **Topic modeling for analyzing online reviews in hotel sector**. *Sci. Tech. Dev. J. - Eco. Law Manag.; 4*(4):1081-1092.