Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
Năm thứ 31, Số 11 (2020), 05–20

www.jabes.ueh.edu.vn

JABES
Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á

Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
http://www.emeraldgrouppublishing.com/services/publishing/jabes/

Phân tích ý kiến khách hàng trong thương mại điện tử – Tiếp cận theo phương pháp học máy kết hợp kiểm định Bootstrap

HỒ TRUNG THÀNH [a,*], TRẦN THỊ ÁNH [a], HUỲNH THANH TUYỀN [a]

[a] *Trường Đại học Kinh tế - Luật, Đại học Quốc Gia TP. Hồ Chí Minh*

THÔNG TIN

Ngày nhận: 10/02/2021
Ngày nhận lại: 28/05/2021
Duyệt đăng: 03/06/2021

Mã phân loại JEL:
C61; C63; C67

Từ khóa:
SOM;
K-Means;
Kiểm định Bootstrap;
Xử lý ngôn ngữ tự nhiên;
Ý kiến khách hàng trực tuyến;
Thương mại điện tử.

Keywords:
SOM;
K-Means;
Bootstrap method;
Natural language

TÓM TẮT

Sự phát triển của thế hệ Web 2.0 đã tạo ra cơ hội tương tác dễ dàng hơn giữa khách hàng và doanh nghiệp thông qua kênh thương mại điện tử. Khách hàng có thể phản hồi ý kiến bằng cách để lại những bình luận dạng văn bản là ngôn ngữ tự nhiên về sản phẩm hay dịch vụ mà họ trải nghiệm. Từ đó, doanh nghiệp có thể quản lý và phân tích ý kiến để hiểu được những trải nghiệm khách hàng nhằm thu hút và giữ chân khách hàng được tốt hơn. Đây là cách tiếp cận quan trọng và hiệu quả để doanh nghiệp có thể tạo được lợi thế cạnh tranh. Trong bài báo này, nhóm tác giả tập trung đề xuất phương pháp phân tích ý kiến khách hàng dựa theo phương pháp xử lý ngôn ngữ tự nhiên kết hợp với phương pháp bản đồ tự tổ chức (SOM) và K-Means. Bên cạnh đó, kỹ thuật kiểm định T với phương pháp Bootstrap được áp dụng để đánh giá kết quả nhằm lựa chọn phương pháp gom cụm phù hợp cho trường hợp dữ liệu là tập văn bản được thu thập từ những phản hồi của khách hàng trên trang thương mại điện tử Tiki.vn. Phương pháp đề xuất có độ chính xác cao và khả năng áp dụng vào phân tích trải nghiệm của khách hàng hiệu quả.

Abstract

The development of the Web 2.0 generation has created an opportunity for interaction between customers and businesses through e-commerce channels more effectively. Customers can post feedbacks by leaving textual comments that are the natural language in Vietnamese related to products or services which they have

* Tác giả liên hệ.
Email: thanhht@uel.edu.vn (Hồ Trung Thành), anhtt@uel.edu.vn (Trần Thị Ánh), tuyenht@uel.edu.vn (Huỳnh Thanh Tuyền).
Trích dẫn bài viết: Hồ Trung Thành, Trần Thị Ánh, & Huỳnh Thanh Tuyền. (2020). Phân tích ý kiến khách hàng trong thương mại điện tử – Tiếp cận theo phương pháp học máy kết hợp kiểm định Bootstrap. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á, 31*(11), 05–20.

processing;
Online customer reviews;
E-commerce.

experienced. As a result, businesses can manage and analyze opinions to deeply understand the customers' experiences to attract and retain their customers. This is an important and effective approach for businesses to create a competitive advantage. In this article, the authors concentrate on proposing a method of analyzing customers' opinions based on the natural language processing method which is combined with Self-Organizing Map (SOM) and K-Means. In addition, the T-test technique with the Bootstrap method is applied to evaluate the results in order to select the appropriate clustering method for the case of a dataset that is collected from customers' feedbacks of Tiki's e-commerce site. The proposed method has high accuracy and effective applicability to customer experience analysis.

1. Giới thiệu

Mục đích quan trọng của doanh nghiệp là tạo ra các sản phẩm dịch vụ thỏa mãn và đáp ứng được đúng nhu cầu của khách hàng; trong đó, phản hồi của khách hàng là thông tin chi tiết về quá trình vận hành sản phẩm hoặc dịch vụ của doanh nghiệp. Doanh nghiệp cần nắm bắt được các thông tin truyền tải trong các phản hồi này và tìm cách cải thiện sao cho khách hàng có những trải nghiệm tốt nhất. Ngoài ra, phản hồi của khách hàng là một trong những nguồn đáng tin cậy có thể được sử dụng để giúp doanh nghiệp kịp thời nắm bắt và xử lý những bất cập của hệ thống nhằm cải thiện chất lượng và hiệu quả hoạt động.

Những năm gần đây, với sự phát triển mạnh mẽ của công nghệ tương tác qua mạng, các trang thương mại điện tử đã tiếp nhận các phản hồi từ khách hàng trực tiếp qua phần bình luận ngay trong sản phẩm. Lượng dữ liệu thu về từ các phản hồi này là rất lớn, do đó, để có thể nắm bắt hết các ý kiến truyền tải bằng phương pháp thủ công là chưa thật khả thi. Với sự hỗ trợ từ các phương pháp học máy cùng các công cụ phân tích, việc phân tích và trích xuất ý nghĩa từ các dữ liệu văn bản (Text) khối lượng lớn này không còn là bất khả thi nữa. Một trong các phương pháp được áp dụng phổ biến là gom cụm và gom nhóm các phản hồi của khách hàng có nội dung tương đồng để từ đó có cơ sở phân hạng cũng như chăm sóc các khách hàng này một cách phù hợp.

Tiki.vn là một trong những trang thương mại điện tử nổi tiếng và uy tín tại Việt Nam. Trang này cho phép người dùng tương tác với sản phẩm qua nhận xét, bình luận từ khách hàng về sản phẩm, trong đó có sách được đăng bán. Các loại sách được kể đến như: Truyện dài, truyện ngắn, sách kinh tế, sách lịch sử, sách chính trị… đều được người dùng chú ý. Ý kiến khách hàng là những phản hồi mà khách hàng cảm nhận được sau khi sử dụng dịch vụ, sản phẩm của doanh nghiệp (Kumar, 2016). Những ý kiến của khách hàng thường bao gồm hai mặt là tích cực và tiêu cực. Tuy nhiên, đôi khi có những ý kiến mang hàm ý cả hai mặt nói trên. Những nhận xét tích cực hay tiêu cực và các chủ đề quan tâm của khách hàng đều sẽ giúp doanh nghiệp biết được ưu, nhược điểm của sách cũng như dịch vụ, từ đó quảng bá, truyền thông hoặc quyết định tiếp tục tái bản hay phát hành sách đó hay không.

Thị trường cạnh tranh ngày một tăng cao, để phục vụ tốt khách hàng thì doanh nghiệp cần phải nắm bắt đúng nhu cầu của họ. Hiểu được những trải nghiệm của khách hàng trên kênh thương mại điện tử sẽ giúp doanh nghiệp phát triển sản phẩm và dịch vụ phù hợp. Tuy nhiên, vấn đề được đưa ra

là làm sao doanh nghiệp có thể biết được khách hàng đang hài lòng hay không hay do thương hiệu đang được nhiều người sử dụng. Để giải quyết bài toán này, cần nghiên cứu và khai thác các bình luận của khách hàng về mặt hàng trên trang thương mại điện tử đó. Nghiên cứu áp dụng các phương pháp học máy không giám sát để phân tích và gom cụm các nhóm bình luận thông qua dữ liệu văn bản để hiểu được những trải nghiệm của khách hàng liên quan đến sản phẩm và dịch vụ được cung cấp trên trang thương mại điện tử. Các bình luận này, trước khi được gom cụm, được tiền xử lý để xây dựng ma trận (không gian véc-tơ) văn bản - từ với tần suất bởi kỹ thuật thống kê TF-IDF (Term Frequency - Inverse Document Frequency). Nghĩa là, phương pháp học máy được áp dụng để gom nhóm cụm từ - cụm từ đại diện cho những bình luận của khách hàng để tìm ra những nội dung mà khách hàng thường trao đổi liên quan đến sản phẩm và dịch vụ của công ty, trong đó bao gồm cả bình luận tích cực hay tiêu cực và các chủ đề quan tâm. Kết quả thực hiện các phương pháp được đánh giá bằng kỹ thuật kiểm định T với phương pháp Bootstrap. Cuối cùng, nghiên cứu ứng dụng công cụ trực quan hoá dữ liệu trên các báo cáo thông minh (Dashboard). Kết quả nghiên cứu giúp các cửa hàng, nhà quản lý doanh nghiệp nắm bắt thông tin dễ dàng và nhanh chóng, từ đó việc phát triển kinh doanh được cải thiện và nâng cao cũng như làm hài lòng và giữ chân khách hàng tốt hơn.

Phần còn lại của bài báo được tổ chức gồm: Phần 2 trình bày các nghiên cứu có áp dụng SOM (Self-Organizing Map) để giải quyết các vấn đề trong kinh doanh; phần 3 trình bày chi tiết cơ sở lý thuyết và ý tưởng của thuật toán SOM; phần 4 đề xuất và thực nghiệm mô hình nghiên cứu, sau đó kiểm định và so sánh kết quả nghiên cứu thu được từ thuật toán SOM với một thuật toán K-Means; cuối cùng, trong phần 5, bài báo sẽ phân tích và trực quan hóa kết quả thu được từ hai thuật toán và thảo luận kết quả.

2. Các nghiên cứu liên quan

Thuật toán SOM được áp dụng khá phổ biến và ứng dụng trong nhiều lĩnh vực. Trong đó, Nakano và cộng sự (2020) đã nghiên cứu và áp dụng mô hình mạng nơ-ron tự tổ chức SOM để dự đoán sự nở hoa của cây anh đào Yoshino; Thanh và cộng sự (2019) nghiên cứu và ứng dụng mạng Kohonen (SOM) kết hợp mô hình chủ đề để gom cụm cộng đồng mạng xã hội có cùng chủ đề quan tâm trong cùng giai đoạn thời gian. Jaye và cộng sự (2019) nghiên cứu những thay đổi trong tương lai về vị trí và tần suất phát sinh xoáy thuận nhiệt đới (Tropical Cyclogenesis) bằng cách áp dụng mô hình mạng nơ-ron tự tổ chức SOM. Trong khi đó, Lin và Li (2009) nghiên cứu dùng thuật toán SOM để gom cụm các nhóm văn bản sau khi dùng HowNet để nhận biết các chữ viết tiếng Trung Quốc. Bên cạnh đó, Liu và cộng sự (2012) nghiên cứu tổng quan và ứng dụng của SOM trong gom cụm văn bản. Phương pháp SOM còn được sử dụng trong các lĩnh vực như: Gom cụm bằng sáng chế (Kohonen và cộng sự, 2000), dịch vụ tài chính (Zhu & Liu, 2014). Ngoài ra, Thanh và Phuc (2015) cũng đã đề xuất phương pháp lai kết hợp giữa mô hình chủ đề và phương pháp Kohonen để gom cụm người dùng dựa trên tập véc-tơ đặc trưng được rút trích từ bình luận trên mạng xã hội.

Gom cụm văn bản với phương pháp SOM có thể được chia thành hai giai đoạn chính: Giai đoạn đầu tiên là tiền xử lý văn bản bao gồm sử dụng mô hình không gian véc-tơ (VSM) để tạo véc-tơ tài liệu đầu ra từ các tài liệu văn bản đầu vào; giai đoạn thứ hai là gom cụm tài liệu áp dụng SOM trên các véc-tơ tài liệu được tạo để thu được các cụm đầu ra (Honkela và cộng sự, 1997; Kaski và cộng sự, 1998).

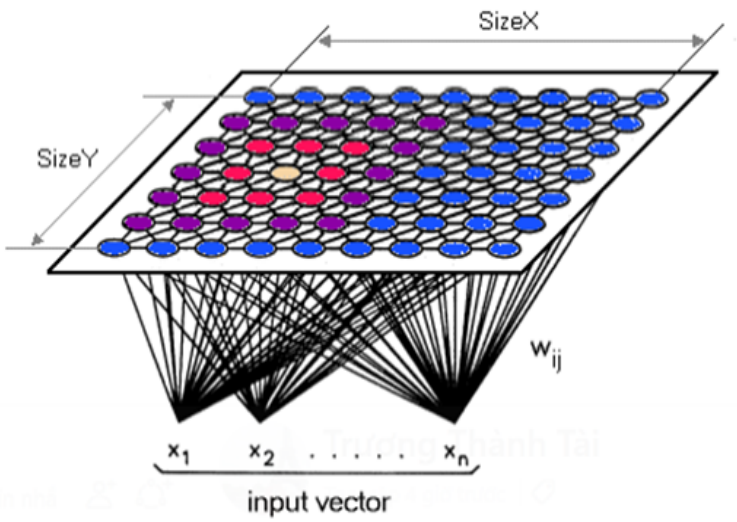
Như đã trình bày ở trên, mạng SOM đã được nghiên cứu và ứng dụng trong nhiều lĩnh vực khác nhau. Tuy nhiên, vẫn còn nhiều tiềm năng nghiên cứu cần được khai thác từ mạng SOM, trong đó, nghiên cứu liên quan đến việc áp dụng mô hình mạng nơ-ron tự tổ chức SOM trong lĩnh vực thương mại điện tử chưa quan tâm nhiều. Cụ thể, với đối tượng nghiên cứu là khách hàng trực tuyến và những bình luận về sản phẩm kinh doanh là sách. Mô hình mạng nơ-ron tự tổ chức SOM được giáo sư Teuvo Kohonen đề xuất đầu tiên vào năm 1982, đặc biệt thích hợp cho việc giảm số chiều của dữ liệu đầu vào, khám phá dữ liệu dạng văn bản và phù hợp với trường hợp nghiên cứu trong bài báo. Bên cạnh đó, thực nghiệm SOM không cần xác định trước số cụm, vì vậy, đây cũng là ưu điểm nổi bật so với các phương pháp khác cùng mục tiêu.

Trong phạm vi bài báo này, mô hình mạng nơ-ron tự tổ chức được ứng dụng để phân tích trường hợp cụ thể là các bình luận của độc giả sau khi mua sách trên trang thương mại điện tử Tiki.vn, với mục tiêu sẽ gom nhóm các bình luận thuộc những cụm tương đồng dựa vào khoảng cách giữa véc-tơ trọng số và véc-tơ huấn luyện từ mẫu đầu vào. Ngoài ra, để đánh giá một cách khách quan, nghiên cứu sẽ so sánh và kiểm định kết quả gom cụm từ SOM với một thuật toán phổ biến khác là K-Means.

3. Cơ sở lý thuyết

3.1. Phương pháp SOM

Bản đồ tự tổ chức (Self-Organizing Map – SOM) là một mạng nơ-ron nhân tạo (Artificial Neural Network – ANN) thuộc nhánh học máy không giám sát (Unsupervised Learning) dựa trên nguyên tắc người thắng được cả (Winner-Takes-All – WTA) (Kohonen, 1995). SOM ban đầu được đề xuất dựa trên ý tưởng rằng các hệ thống có thể được thiết kế để mô phỏng sự hợp tác tập thể của các tế bào thần kinh trong não người. Đây là một phương pháp học máy không giám sát được sử dụng rộng rãi trong khai thác dữ liệu, trực quan hóa dữ liệu phức tạp, xử lý hình ảnh, nhận dạng giọng nói, kiểm soát quá trình, chuẩn đoán trong công nghiệp và y học, và xử lý ngôn ngữ tự nhiên (Kohonen, 2013). Mô hình SOM điển hình được tạo thành từ véc-tơ của các nơ-ron đầu vào (lớp nơ-ron đầu vào), một mảng các nút xem như là lớp đầu ra (bản đồ Kohonen) và một ma trận các kết nối giữa mỗi nơ-ron đầu ra với tất cả các lớp nơ-ron đầu vào. Mỗi liên kết giữa các nơ-ron đầu vào và nơ-ron đầu ra của bản đồ Kohonen tương ứng với một véc-tơ trọng số, kích thước (số chiều) của véc-tơ trọng số này bằng kích thước của nơ-ron đầu vào. Nói cách khác, mỗi nơ-ron của bản đồ Kohonen sẽ có thêm một véc-tơ trọng số n chiều (với n là số chiều trong mẫu đầu vào).



Hình 1. Ma trận trọng số (Weight Matrix) kết nối giữa Vector nhập và các Neurons trên lớp ra Kohonen

Nguồn: Ettaouil và cộng sự (2011).

Nói một cách đơn giản hơn, SOM có thể chuyển đổi một không gian nhiều chiều trong mẫu dữ liệu đầu vào thành một không gian ít chiều hơn (thường là hai chiều) và vì thế nó cũng được xem như là một phương pháp giảm số chiều dữ liệu đầu vào.

Đối với bản đồ Kohonen hai chiều, các nơ-ron nằm trên bản đồ có thể tồn tại hai loại cấu trúc liên kết là cấu trúc hình lục giác hoặc hình chữ nhật. Tuy nhiên, cấu trúc liên kết hình lục giác đều thì tốt hơn trong tác vụ trực quan hoá vì mỗi nơ-ron sẽ có 6 nơ-ron lân cận trong khi với cấu trúc hình chữ nhật thì chỉ là 4 nơ-ron.

3.2. Phương pháp học ganh đua

Học ganh đua (Competitive Learning) là một trong những thuật toán thuộc nhánh học máy không giám sát liên quan đến việc dùng các phương pháp quy nạp để phát hiện tính quy chuẩn được thể hiện trong tập dữ liệu (Rumelhart & Zipser, 1985).

Mô hình SOM học bằng cách thông qua việc tự tổ chức các nơ-ron ngẫu nhiên của bản đồ Kohonen liên kết với các lớp nơ-ron đầu vào qua véc-tơ trọng số. Các véc-tơ trọng số này được cập nhật ở mọi thời điểm trong quá trình học. Sự thay đổi trọng số phụ thuộc vào sự giống nhau hoặc tương đồng về không gian giữa mẫu các lớp nơ-ron đầu vào và mẫu các nơ-rơn thuộc bản đồ Kohonen (Michalski và cộng sự, 1999).

Gọi $x_i \in \mathbb{R}M$ và $1 \leq i \leq N$ là véc-tơ huấn luyện cho bản đồ Kohonen được chọn ngẫu nhiên đại diện cho tài liệu i trong tập ngữ liệu (Corpus), trong đó, M là số lượng đặc tính được lập chỉ mục (Index) và N là số tài liệu trong tập dữ liệu đầu vào. Mỗi nơ-ron trong bản đồ Kohonen có M kết nối. Gọi véc-tơ trọng số là w_j, véc-tơ này được liên kết với nơ-ron thứ j trong bản đồ Kohonen có nơ-ron J, trong đó:

$$w_j = \{w_{jm} \mid 1 \leq m \leq M\} \text{ và } 1 \leq j \leq J \quad (1)$$

Thuật toán học ganh đua trong SOM bao gồm các bước sau:

- *Bước 1*: Khởi tạo các véc-tơ trọng số của các nơ-ron trong bản đồ Kohonen.

- *Bước 2*: Chọn ngẫu nhiên một véc-tơ huấn luyện x_i trong tập ngữ liệu đầu vào.

Với mỗi véc-tơ x_i trong tập ngữ liệu đầu vào, véc-tơ trọng số w_j tương ứng cũng có số chiều tương tự.

- *Bước 3*: Xác định nơ-ron c còn được gọi là BMU (Best Matching Unit) hay nơ-ron chiến thắng.

Nơ-ron c là nơ-ron thuộc bản đồ Kohonen và có khoảng cách Euclid (Euclidean Distance) nhỏ nhất tính từ véc-tơ x_i được chọn ngẫu nhiên từ bước 2 đến véc-tơ trọng số wj của chính nó. Khi đó, nơ-ron c được gọi là BMU (Best Matching Unit – tạm dịch là: Đơn vị khớp nhất).

Công thức (2) tính khoảng cách Euclid giữa véc-tơ x_i và véc-tơ trọng số w_j và chọn BMU có thể được viết lại như sau:

$$c: ||x_i(t) - w_c(t)|| = \min (||x_i(t) - w_j(t)||) \quad (2)$$

Ký hiệu thời gian rời rạc t thể hiện số thứ tự của vòng lặp trong thuật toán.

- *Bước 4*: Cập nhật véc-tơ trọng số của BMU và các véc-tơ trọng số của các nơ-ron xung quanh nó để di chuyển BMU đến gần hơn so với véc-tơ đầu vào.

Bản đồ Kohonen sẽ tự thích ứng với mẫu đầu vào trong tất cả các lần lặp lại của quá trình học và việc tự thích ứng được thực hiện bằng cách liên tục giảm dần khoảng cách Euclid giữa véc-tơ x_i trong Bước 2 với véc-tơ trọng số của BMU. Tham số tần suất học α(t) kiểm soát mức độ thích ứng liên tục giảm dần trong quá trình học.

Hàm lân cận được sử dụng để xác định phạm vi không gian giữa BMU và các nơ-ron xung quanh nó. Bán kính của vùng lân cận với BMU là tâm, được tính bằng khoảng cách giữa BMU đến các nơ-ron còn lại. Điều này đảm bảo rằng các nơ-ron trong vùng lân cận có sự tự điều chỉnh tốt hơn các nơ-ron khác nằm ngoài vùng. Hàm Gaussian (Hàm số bậc 4) có thể đóng vai trò là hàm lân cận với r_i đại diện cho một véc-tơ hai chiều chỉ đến vị trí của nơ-ron *j* nằm trong bản đồ Kohonen, và $||r_j - r_c||$ thể hiện khoảng cách từ nơ-ron *j* đến BMU. Công thức (3) trình bày phương pháp tính bán kính vùng lân cận với tâm là BMU như sau:

$$h_{c(x_i),j} = e^{-\left\|r_j - r_c\right\|^2 / 2\sigma^2(t)} \quad (3)$$

Tham số σ thay đổi theo thời gian thể hiện sự thay đổi về độ lớn của bản đồ Kohonen trong quá trình tự điều chỉnh, theo đó, một vùng diện tích lớn của bản đồ Kohonen được xem như đối tượng của sự tự điều chỉnh khi bắt đầu quá trình học. Số lượng nơ-ron lân cận cũng bị ảnh hưởng bởi sự tự điều chỉnh, chúng sẽ giảm dần trong suốt quá trình học.

Quy tắc học cho bất kỳ nơ-ron *j* nào thuộc bản đồ Kohonen có thể được biểu diễn như công thức. Công thức tính số nơ-ron trong vùng lân cận với số vòng lập t như sau:

$$w_j(t + 1) = w_j(t) + \alpha(t) h_{c(x_i),j}(t) \, . \, [x_i(t) - w_j(t)] \quad (4)$$

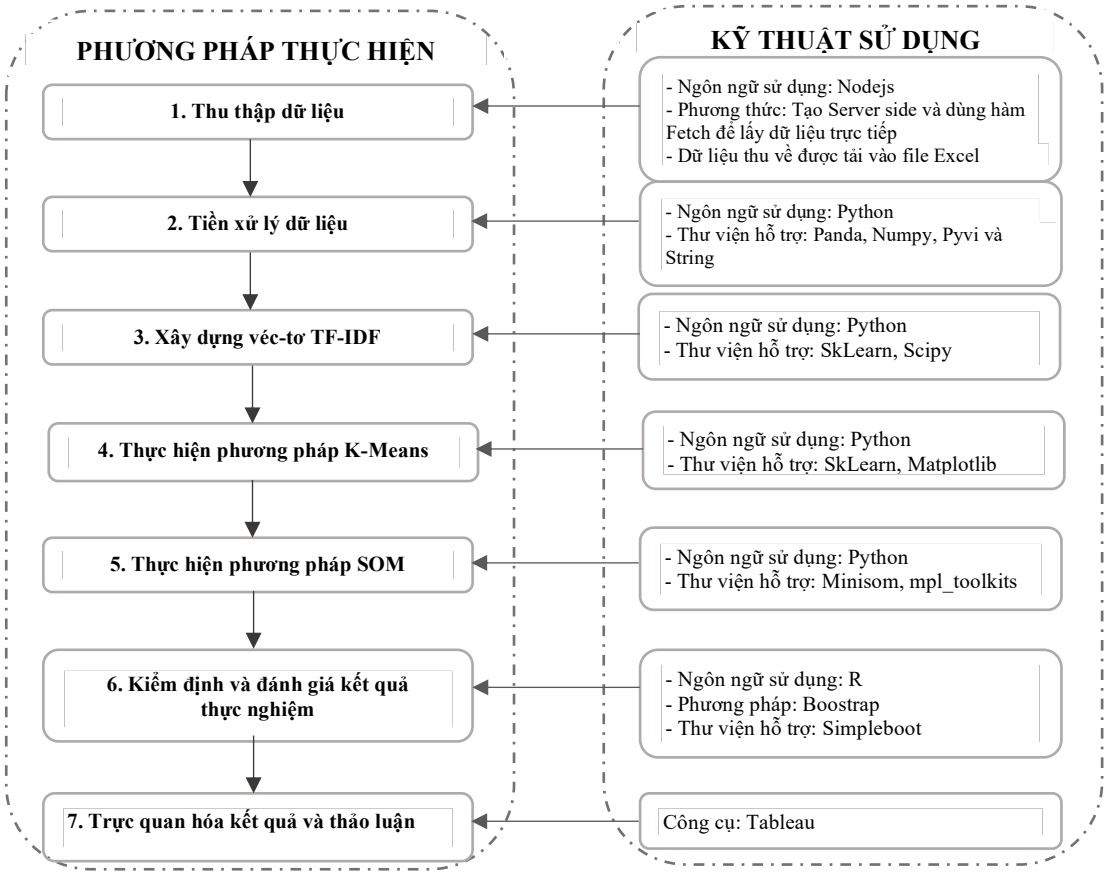
Bước 5: Tăng số lần lặp t của vòng lặp. Khi t đạt đến giá trị tối đa T đã thiết lập trước đó, quá trình học sẽ kế thúc; Nếu không thì giảm α(t) ($0 < \alpha(t) < 1$) và kích thước vùng lân cận rồi quay lại Bước 2.

SOM có hai ưu điểm so với các phương pháp gom cụm khác là phép chiếu phi tuyến của không gian đầu vào và bảo toàn cấu trúc liên kết cụm. Do đó, SOM có thể trích xuất các mối quan hệ phi tuyến vốn có giữa các tài liệu và các cụm tài liệu tương đồng được ánh xạ gần nhau, và từ đó tìm ra lớp cấu trúc cao hơn (cấu trúc nổi) (Ultsch & Herrmenn, 2005); trong bản đồ tự tổ chức cải tiến

(Evolve Self-Organizing Maps – ESOM), ranh giới giữa các cụm là "không rõ ràng", mức độ phân tách giữa các "vùng" trong bản đồ Kohonen (tức là các cụm) được mô tả bằng "độ dốc" (Ultsch & Herrmenn, 2005); bên cạnh đó, với phương pháp gom cụm SOM không cần xác định số cụm K. Với những ưu điểm hiệu quả trên, mô hình SOM đã thúc đẩy các nhà nghiên cứu sử dụng nó để gom cụm văn bản và trực quan hóa (Yen & Wu, 2008).

3.3. Mô hình nghiên cứu và phương pháp thực hiện

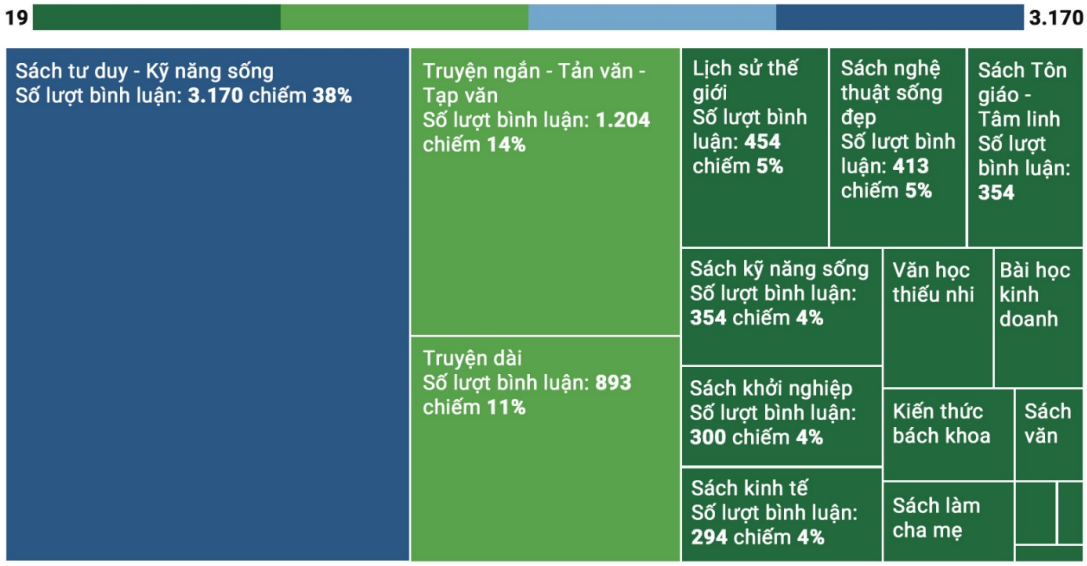
3.3.1. Mô hình nghiên cứu



Hình 2. Mô hình nghiên cứu và phương pháp thực hiện

3.3.2. Mô tả Bộ dữ liệu

Trước tiên, dữ liệu được thu thập từ các bình luận, đánh giá của khách hàng mua sách từ trang thương mại điện tử Tiki.vn đối với các đầu sách tiếng Việt. Bên cạnh những bình luận của khách hàng, nhóm tác giả còn thu thập thêm các thông tin khác có liên quan như: Tên tài khoản của khách hàng, tên sách được nhận xét, và thể loại sách mà khách hàng mua. Những dữ liệu bao gồm các bình luận có thông tin của người nhận xét, có điểm đánh giá đối với đầu sách mà người dùng mua và có thông tin thời gian viết bình luận. Kết quả sau quá trình thu thập dữ liệu, bộ dữ liệu gồm có 8.394 quan sát với 17 nhóm nội dung sản phẩm như sau:



Hình 3. Số lượt bình luận phân theo nhóm sản phẩm

Hình 3 thể hiện tỷ lệ bình luận được thu thập. Sách Tư duy – Kỹ năng sống có 3.170 lượt bình luận (tương ứng 38%), Truyện ngắn – Tản văn – Tạp văn: 1.204 lượt bình luận (14%), Truyện dài: 893 lượt bình luận (11%), Lịch sử thế giới: 454 lượt bình luận (5%), sách Nghệ thuật sống đẹp: 413 lượt bình luận (5%), sách Kỹ năng sống và sách Tôn giáo – Tâm linh: cùng 354 lượt bình luận (4%), sách Khởi nghiệp: 300 lượt bình luận (4%), sách Kinh tế: 294 lượt bình luận (4%). Còn lại, các đầu sách khác với số lượng bình luận không quá nổi trội.

3.3.3. Tiền xử lý dữ liệu

Phương pháp tiền xử lý được tiến hành dựa theo phương pháp xử lý ngôn ngữ tự nhiên theo thứ tự các bước sau:

- *Bước 1:* Sử dụng các thư viện liên quan: Pandas (đọc tập dữ liệu), Numpy (đưa dữ liệu vào dạng véc-tơ), Matplotlib (trực quan hóa dữ liệu), Sklearn (thư viện các thuật toán học máy của Pyhton), Pyvi (tách từ tiếng Việt), Scipy (cung cấp khá nhiều Module tính toán), String (dùng để tách các dấu câu).

- *Bước 2:* Tách dấu câu: Dùng thư viện hỗ trợ String để tạo định nghĩa hàm tách dấu câu.

- *Bước 3:* Sau khi tách dấu câu, sử dụng thư viện hỗ trợ Pyvi để tách từ dựa vào bộ từ điển có sẵn trong thư viện. Thuật toán thực hiện với từng đối tượng, cần xây dựng vòng lặp với số lần lặp bằng đúng số đối tượng để có thể tách toàn bộ các từ trong toàn bộ đối tượng.

- *Bước 4:* Loại bỏ từ dừng (Stopwords) với bộ từ điển từ dừng có sẵn. Sau đó, sử dụng phương pháp TF-IDF để véc-tơ hóa các bình luận dạng văn bản, từ đó có thể áp dụng các thuật toán máy học đã được cài đặt trong Sklearn và Minisom (một thư viện triển khai SOM mức tối giản).

- *Bước 5:* Ngoài ra, để tăng độ chính xác sau khi lọc từ, véc-tơ TF-IDF được lọc các từ sai chính tả, viết tắt, từ không có nghĩa… bằng phương pháp bán thủ công và sau đó được so khớp với bộ từ

điển để tạo ra véc-tơ kết quả là ma trận TF-IDF tốt hơn. Kết quả loại bỏ hơn 1.700 từ không có nghĩa trong trường hợp dữ liệu của nghiên cứu này.

- *Bước 6:* Tạo lập ma trận TF-IDF cuối cùng với dòng là thứ tự các bình luận (văn bản – Doc), cột là các từ sau xử lý. Cụ thể trong trường hợp này, số cột của ma trận TF-IDF tương ứng là 5.434 từ (đã lọc trùng) trong tập dữ liệu.

Bảng 1.

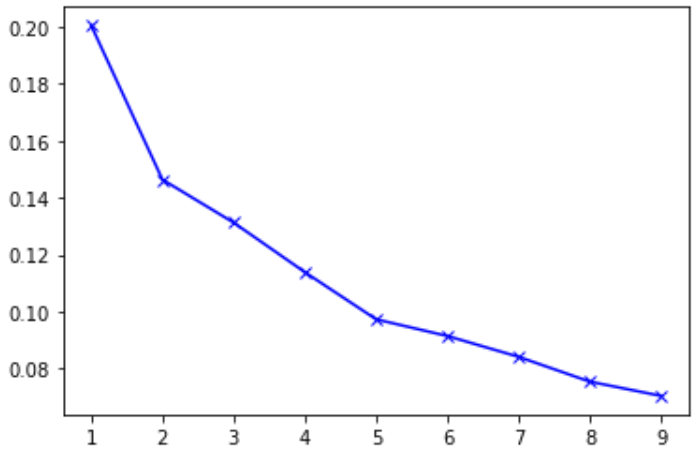
Một số mẫu đại diện kết quả ma trận TF-IDF

Văn bản (bình luận)	Từ (Word)						
	cuồng_phong	cuộc_sống	cuộc_đời	cánh	câu	câu_chuyện	cười
Doc_5	0,000	0,000	0,106	0,000	0,000	0,000	0,000
Doc_15	0,102	0,056	0,000	0,188	0,000	0,000	0,000
Doc_20	0,000	0,000	0,000	0,000	0,162	0,000	0,000
Docx_23	0,136	0,000	0,000	0,125	0,000	0,000	0,000
Doc_25	0,000	0,150	0,000	0,000	0,000	0,000	0,000
Doc_411	0,000	0,000	0,170	0,000	0,000	0,000	0,000
Doc_412	0,000	0,000	0,000	0,000	0,000	0,089	0,195
Doc_874	0,000	0,000	0,000	0,000	0,177	0,000	0,000

3.3.4. Thực nghiệm phương pháp và mô hình nghiên cứu

Bên cạnh ứng dụng SOM vào thực nghiêm mô hình, nghiên cứu còn thực hiện gom cụm dữ liệu với giải thuật K-Means để so sánh kết quả, đồng thời sử dụng phương pháp phân tích thành phần chính PCA (Principal Component Analysis) sử dụng phép biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn nhằm tối ưu hóa việc thể hiện sự biến thiên của dữ liệu (Jolliffe & Cadima, 2016) được tích hợp trong thư viện Sklearn để giảm số chiều dữ liệu ban đầu vào xuống còn hai chiều với mục tiêu tăng tốc độ tính toán và hướng đến trực quan hóa kết quả gom cụm. Hàm PCA tiến hành xây dựng một không gian mới ít chiều hơn nhưng lại có khả năng biểu diễn dữ liệu tốt tương đương không gian ban đầu, nghĩa là vẫn bảo đảm được độ biến thiên (Variability) của dữ liệu trên mỗi chiều mới.

Trường hợp với K-Means, sử dụng phương pháp Elbow để tìm số cụm k phù hợp bằng cách trực quan đồ thị. Phương pháp Elbow được sử dụng để xác định số lượng cụm tối ưu trong gom cụm K-Means. Phương pháp Elbow vẽ biểu đồ giá trị của hàm chi phí được tạo ra bởi các giá trị khác nhau của k cụm (Syakur và cộng sự, 2018). Nếu k tăng, độ nghiêng trung bình sẽ giảm, mỗi cụm sẽ có ít cá thể cấu thành hơn và các thể hiện sẽ gần với trung tâm tương ứng của chúng hơn. Tuy nhiên, sự cải thiện về độ nghiêng trung bình sẽ giảm khi k tăng. Giá trị của k mà tại đó sự cải thiện về độ nghiêng giảm nhiều nhất được gọi là giá trị khuỷu, tại đó, chúng ta nên dừng việc chia dữ liệu thành các cụm xa hơn.



Hình 4. Đồ thị kết quả hệ số k cụm từ thực nghiệm dữ liệu với phương pháp Elbow

Theo kết quả ở Hình 4, chọn số cụm k = 3, chạy thuật toán K-Means, kết quả trả về trong cột dữ liệu mới đặt tên là K_mean_group với giá trị là các cụm xuất ra từ thuật toán.

Trường hợp với SOM, nghiên cứu sử dụng thư viện Minisom trong Python để tiến hành thuật toán SOM với kết quả số cụm là 3 và số lần lặp là 500 lần, sau đó tạo thêm cột dữ liệu mới, đặt tên là "SOM-group" với giá trị là kết quả trả về từ thuật toán SOM.

4. Kiểm định phương pháp và đánh giá kết quả

Bộ dữ liệu ban đầu đã được bổ sung vào hai cột dữ liệu: (1) K_mean_group: Giá trị là các cụm thu được từ thuật toán K-Means, được đánh số từ 0 cho đến 2; (2) SOM_group: Giá trị là các cụm thu được từ thuật toán SOM, được đánh số từ 0 cho đến 2.

Từ hai cột dữ liệu mới này, kết quả sẽ được xem xét trả lời câu hỏi sau: Liệu rằng các cụm được phân bởi thuật toán K-Means và SOM có khác biệt nhau hay không và độ chính xác của từng thuật toán như thế nào? Nếu có, liệu rằng sự khác biệt này có ý nghĩa thống kê hay không? Phương pháp nào được lựa chọn? Để trả lời các câu hỏi trên, kỹ thuật kiểm định T với phương pháp Bootstrap được nghiên cứu và áp dụng với số lần lặp là 1.000 lần và cho kết quả như trên đoạn mã sau:

```
> quantile(diff, probs = c(0.025, 0.5 , 0.975))
    2.5%      50%    97.5%
1.325950 1.438000 1.540025
> diff1 <- X - Y
> t.test(diff1)

        One Sample t-test

data:  diff1
t = 63.84, df = 8393, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.196318 1.272112
sample estimates:
mean of x
 1.234215
```

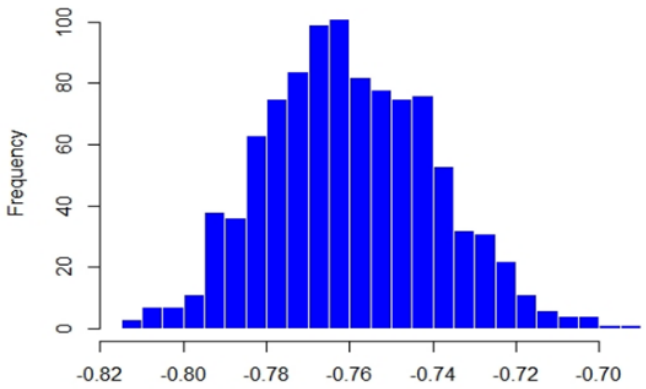
Kết quả được trình bày tại Bảng 2 dưới đây.

Bảng 2.
Phân bố giá trị trong khoảng tin cậy với kỹ thuật kiểm định T và phương pháp Bootstrap

	Giá trị 1	Giá trị 2	Giá trị 3
Phân vị (Quantile)	2,5%	50,0%	97,5%
Độ chênh lệch	–0,795	–0,760	–0,719

Về ý nghĩa, kết quả gom cụm từ hai thuật toán (K-Means và SOM) đối với bộ dữ liệu thu được có khác nhau. Tuy nhiên, như đã được kiểm định T bằng R với phương pháp Bootstrap trên, độ chênh lệch không quá lớn (–0,760) và điều này được đề cập đến độ chênh lệch số thứ tự nhãn của mỗi cụm (khi đề cập đến sự tương đồng kết quả số bình luận và từ - cụm từ trong từng nhãn cụm) của cả hai phương pháp (nội dung này được thảo luận chi tiết trong mục 5 của bài báo). Sự khác nhau này hoàn toàn có ý nghĩa thống kê. Bằng chứng là trị số p-value (mức ý nghĩa) nhỏ hơn 0,05, hiệu số sự khác biệt giữa hai nhóm có khoảng tin cậy 95% từ –0,795 đến –0,719.

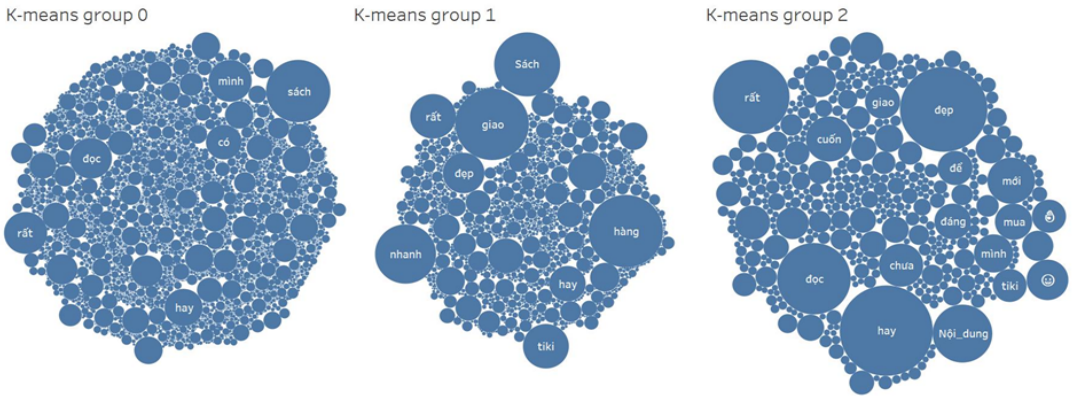
Trực quan hóa sự phân phối khác biệt giữa hai cụm sử dụng hàm Hist trong R:



Hình 5. Phân bố độ chênh lệch của giá trị trung bình giữa hai cụm

5. Trực quan hóa và thảo luận kết quả

Để có thể dễ dàng phân tích được kết quả gom cụm, nghiên cứu sử dụng giải thuật SOM để ước lượng số cụm cần được tách ra từ tập dữ liệu trên, sau đó, dựa trên mô hình không gian véc-tơ, cụ thể là TF-IDF, cuối cùng là sử dụng thuật toán K-Means để tách tập dữ liệu thành ba cụm như lựa chọn ban đầu. Song song đó, mỗi từ được tách ra sẽ gán với mã cụm đã được phân cho bình luận đó. Thực nghiệm với bộ dữ liệu đầu vào, kết quả thu được 170.000 từ được gắn nhãn theo từng cụm tương ứng. Sau đó, kết quả gom cụm được trực quan hóa thông qua những biểu đồ thể hiện từ trong mỗi cụm. Mức độ lớn hay nhỏ trong từng cụm thể hiện tần suất của từ trong tập dữ liệu văn bản từ bình luận (Hình 6).



Hình 6. Nội dung từng cụm (Group) với thuật toán K-Means

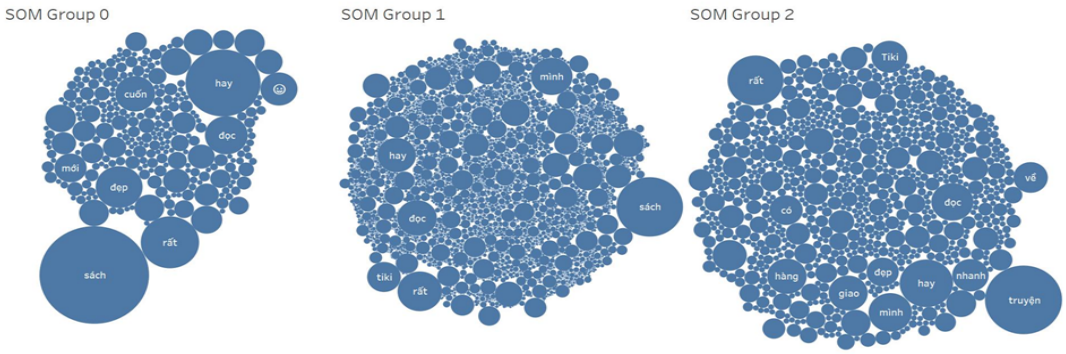
Kết quả Hình 6 với từng chủ đề theo từng cụm, nội dung từng cụm đã được gom cụm bởi thuật toán K-Means và có ý nghĩa được mô tả trong Bảng 3.

Bảng 3.

Ý nghĩa của từng cụm theo thuật toán K-Means

Nhãn cụm	Nội dung
Cụm 0	Chủ yếu là phản hồi tích cực, nói về nội dung tác phẩm
Cụm 1	Vẫn là những phản hồi tích cực về dịch vụ của Tiki (như dịch vụ giao hàng)
Cụm 2	Gồm những phản hồi tích cực, nhưng nói một cách chung chung

Trường hợp đối với mạng SOM, kết quả thử nghiệm như sau:



Hình 7. Nội dung từng cụm với thuật toán SOM

Tương tự, kết quả như Hình 7 trên với từng chủ đề theo từng cụm, nội dung từng cụm đã được gom cụm bởi thuật toán SOM và có ý nghĩa được mô tả trong Bảng 4.

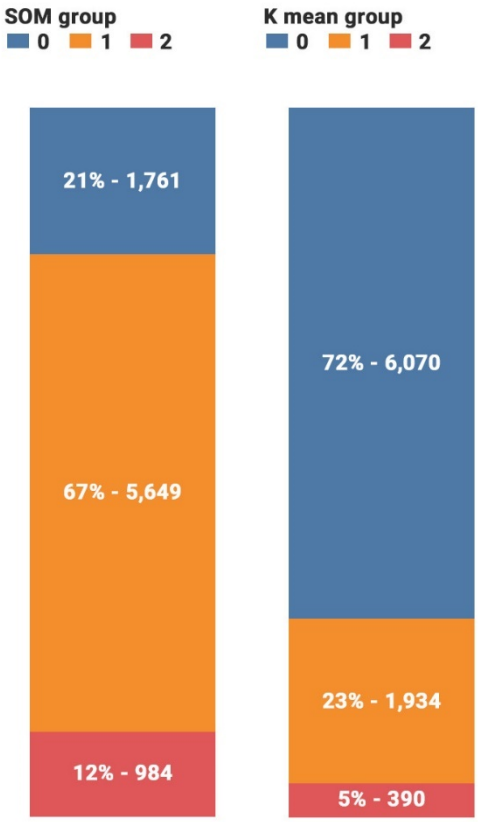
Bảng 4.

Ý nghĩa từng cụm theo thuật toán SOM

Nhãn cụm	Nội dung
Cụm 0	Chủ yếu là phản hồi tích cực, nhưng nói một cách chung chung
Cụm 1	Chủ yếu là phản hồi tích cực, nói về nội dung tác phẩm
Cụm 2	Gồm những phản hồi tích cực, nhưng nội dung về các dịch vụ của Tiki

Ngoài nội dung, nghiên cứu còn xem xét khía cạnh số lượng các bình luận trong mỗi nhóm của mỗi thuật toán và tóm tắt lại như biểu đồ được trình bày trong Hình 8.

Qua biểu đồ trong Hình 8, có thể thấy rằng giữa số lượng bình luận được gom trong mỗi nhãn nhóm (Group) ứng với từng thuật toán có sự khác biệt, nhưng sự khác biệt về số lượng này không quá lớn khi xét đến nhãn các cụm có nội dung tương đồng nhau, phù hợp với kết quả trả về sau khi kiểm định bằng kỹ thuật kiểm định T trên R. Ngoài ra, qua số liệu trên, nhóm tác giả thấy rằng, đối với thuật toán SOM, nhóm 1 chiếm tỷ lệ cao nhất (khoảng 67%), đối với thuật toán K-Means, nhóm 0 lại chiếm tỷ lệ cao nhất (khoảng 70%). Kết quả này khớp với nội dung từng nhóm mà nghiên cứu đã nhận định ở phần trên: Nhóm có nhãn là 1 từ kết quả của thuật toán SOM tương đồng với nhóm có nhãn là 0 từ kết quả của thuật toán K-Means.



Hình 8. Số lượng bình luận và tỷ trọng sau khi gom cụm với từng thuật toán

Với kết quả gom cụm cũng như kết quả kiểm định T với phương pháp Bootstrap trên cho thấy trường hợp dữ liệu của nghiên cứu có thể sử dụng cả hai phương pháp gom cụm trên. Tuy nhiên, với những ưu điểm của mạng SOM về việc không cần xác định trước số cụm và khả năng biểu diễn trực quan kết quả khám phá cộng đồng trên lớp ra Kohonen hai chiều, nghiên cứu lựa chọn SOM là phương pháp phù hợp để phân tích cho trường hợp khi bổ sung dữ liệu học đầu vào tương ứng trong nghiên cứu.

Đối với thuật toán K-Means, việc xác định trước hệ số K cụm chính là vấn đề đặt ra nếu dữ liệu ngày càng lớn và mở rộng phạm vi chủ đề. Trong khi việc ước lượng hệ số K này cũng ở mức tương đối mặc dù có phương pháp Elbow hỗ trợ. Bên cạnh đó, trong nghiên cứu này, khi không gian véc-tơ với số chiều càng lớn, việc giảm số chiều để tăng hiệu suất tính toán, lưu trữ và trực quan hóa kết quả cũng là một hạn chế đối với giải thuật K-Means. Việc này dẫn đến tiêu tốn nguồn tài nguyên hơn khi phải áp dụng thêm hàm PCA để giảm bớt số chiều trước khi gom cụm.

6. Kết luận và hướng phát triển

6.1. Kết luận

Kết quả nghiên cứu đóng góp ba nội dung chính:

- *Thứ nhất*, kết quả nghiên cứu đã xây dựng bộ từ điển "từ dừng" để lọc từ, bộ này được nâng cấp theo từng bộ dữ liệu đầu vào. Càng nhiều bộ dữ liệu được sử dụng, bộ từ điển này sẽ ngày càng phong phú và sẽ giúp tăng độ chính xác của dữ liệu đầu ra trong quá trình tiền xử lý dữ liệu.

- *Thứ hai*, ứng dụng phương pháp gom cụm không giám sát gồm K-Means và SOM để thực nghiệm trên bộ dữ liệu thực tế được thu thập từ trang thương mại điện tử và đánh giá kết quả.

- *Thứ ba*, sử dụng kỹ thuật kiểm định T với phương pháp Bootstrap để so sánh kết quả của hai phương pháp.

Và cuối cùng, kết quả nghiên cứu được trực quan hóa, phân tích và thảo luận những ưu điểm, hạn chế và từ đó lựa chọn phương pháp nghiên cứu phù hợp.

6.2. Hướng phát triển

Trong phạm vi của nghiên cứu, các bình luận từ ý kiến của khách hàng được khai thác bao gồm cả bình luận tích cực và tiêu cực và các chủ đề quan tâm được phân tích và xem xét dựa trên dữ liệu thu thập. Vì vậy, nghiên cứu tiếp theo sẽ được mở rộng xem xét đến nhóm bình luận trung tính từ dữ liệu thu thập kết hợp với tỷ lệ đánh giá (Rating) tương ứng với mỗi bình luận. Nghiên cứu cũng sẽ tích hợp phương pháp phân tích quan điểm và cảm xúc (Sentiment và Emotion Analysis) vào mô hình được đề xuất để kết quả nghiên cứu có thêm những góc nhìn đa chiều khác nhau về những trải nghiệm của khách hàng sau khi sử dụng sản phẩm và dịch vụ. Hơn thế nữa, nghiên cứu sẽ thu thập thêm dữ liệu từ nhiều nguồn khác nhau có liên quan, sử dụng thêm phương pháp gom cụm K-Medoids, và bổ sung thêm các phương pháp kiểm định khác để đánh giá kết quả từ các phương pháp gom cụm.

Bên cạnh đó, nghiên cứu tiếp theo sẽ kết hợp với mô hình chủ đề có yếu tố thời gian để gom cụm từng chủ đề mà khách hàng quan tâm và nghiên cứu phương pháp gán nhãn tự động với những chủ đề khám phá (các cụm) được từ tập dữ liệu thu thập. Từ đó, giúp cho công ty cũng như nhà quản lý

hiểu rõ hơn về những vấn đề mà khách hàng thường xuyên quan tâm nhằm cải thiện hiệu quả kinh doanh và có chính sách chăm sóc khách hàng hiệu quả hơn■

Lời cảm ơn

Nhóm tác giả xin gửi lời cảm ơn đến các anh/chị thuộc nhóm nghiên cứu do TS. Hồ Trung Thành hướng dẫn đang là học viên cao học tại Khoa Công nghệ Thông tin Kinh doanh, Trường Đại học Kinh tế TP. Hồ Chí Minh đã hỗ trợ để nhóm tác giả có thể hoàn thành tốt nghiên cứu.

Tài liệu tham khảo

Jaye, A. B., Bruyère, C. L., & Done, J. M. (2019). Understanding future changes in tropical cyclogenesis using Self-Organizing Maps. *Weather and Climate Extremes, 26*, 100235. doi: 10.1016/j.wace.2019.100235

Ettaouil, M., Lazaar, M., Elmoutaouakil, K., & Glanou, Y. (2011). A new architecture optimization model for the Kohonen networks and clustering. *Journal of Advanced Research, 3*(1), 14–32.

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1997). Websom-self-organizing maps of document collections. *Proceedings of the Work shop on Self-Organizing Maps (WSOM '97)* (pp. 310–315), Espoo, Finland.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical transactions of the Royal society A Mathematical Physical and Engineering Sciences, 374*(2065). doi: 10.1098/rsta.2015.0202

Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1998). Websom-self-organizing maps of document collections. *Neurocomputing, 21*(1–3), 101–117.

Kohonen, T. (1995). *Self-Organizing Maps.* Berlin: Spinger-Verlag.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, *11*(3), 574–585.

Kumar, V. (2016). Introduction: Is Customer Satisfaction (Ir)relevant as a Metric?. *Journal of Marketing, 80*(5), 108–109. doi:10.1509/jm.80.5.1

Lin, G., & Li, S. (2009). Clustering method using hypergraph models based on Set Pair Analysis. *2009 IEEE International Symposium on IT in Medicine & Education* (1194–1197). doi: 10.1109/ITIME.2009.5236279

Liu, Y.-C., Liu, M., & Wang, X.-L. (2012). Application of self-organizing maps in text clustering: A review. In M. Johnsson (Ed.), *Applications of Self-Organizing Maps* (Chapter 10). Rijeka: InTech.

Michalski, R. S., Bratko, I., & Kubat, M. (1999). *Machine Learning And Data Mining Methods And Applications*. Wiley

Nakano, T., Nagai, S., Yamatogi, T., Kunhara, T., & Okamura, K. (2020). Use of sea surface discoloration to monitor and discriminate the causative genera of harmful algal blooms (HABs):

Practical use of digital repeat photography. *Ecological Informatics, 59*. doi: /10.1016/j.ecoinf.2020.101114

Ultsch, A., & Herrmann, L. (2005). *The Architecture of Emergent Self-Organizing Maps to Reduce Projection Errors*. ESANN 2005, 13th European Symposium on Artificial Neural Networks, Bruges, Belgium.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, *9*, 75–112.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, 336(1). Retrieved from https://iopscience.iop.org/article/10.1088/1757-899X/336/1/012017

Till, B. C., Longo, J., Dobell, A. R., & Driessen, P. F. (2014). Self-organizing maps for latent semantic analysis of free-form text in support of public policy analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(1), 71–86.

Thanh, H., & Phuc, D. (2015). Discovering Communities of Users on Social Networks Based on Topic Model Combined with Kohonen Network. *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)* (pp. 268-273). doi: 10.1109/KSE.2015.54

Thanh, H. T., Hung, N. Q., & Thanh, T. D. (2019). Applying topic model combined with Kohonen networks to discover and visualize communities on social networks. *Science & Technology Development Journal-Economics-Law and Management*, *3*(3), 311–326.

Yen, G. G., & Wu, Z. (2008). Ranked Centroid Projection: A Data Visualization Approach with Self-Organizing Maps. *IEEE Transactions on Neural Networks, 19*(2), 245–259.

Zhu, J., & Liu, S. (2014). SOM network-based clustering analysis of real estate enterprises. *American Journal of Industrial and Business Management*, *4*(3), 167–173.