Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á
Năm thứ 33, Số 12 (2022), 21–38

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/

Mô hình hệ thống khuyến nghị sản phẩm trong kinh doanh trực tuyến dựa vào khai thác dữ liệu phi cấu trúc

LÊ TRIỆU TUẤN [a,*], PHẠM MINH HOÀN [b]

[a] *Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên*

[b] *Trường Đại học Kinh tế Quốc dân*

THÔNG TIN

Ngày nhận: 12/09/2022
Ngày nhận lại: 27/11/2022
Duyệt đăng: 28/11/2022

Mã phân loại JEL:
C61; C63; C67.

Từ khóa:
Hệ thống khuyến nghị trực tuyến;
Xếp hạng ảo;
Xếp hạng thực;
Khai thác dữ liệu phi cấu trúc.

Keywords:
Online Recommendation System;
Virtual rating;
Real rating;

TÓM TẮT

Dữ liệu phi cấu trúc được tạo ra bởi khách hàng trên các trang thương mại điện tử ngày càng nhiều và trở nên quan trọng để nghiên cứu và phát triển hệ thống khuyến nghị sản phẩm, hỗ trợ khách hàng trực tuyến ra quyết định mua hàng. Những dạng dữ liệu này đang cung cấp hướng nghiên cứu mới để giải quyết các vấn đề thách thức của hệ thống khuyến nghị truyền thống. Bài báo đề xuất mô hình hệ thống khuyến nghị sản phẩm trực tuyến dựa vào khai thác các bình luận dưới dạng văn bản của khách hàng. Hệ thống bao gồm hai phân hệ: Phân hệ thứ nhất là quy trình khai thác dữ liệu phi cấu trúc; phân hệ thứ hai thực hiện khuyến nghị sản phẩm. Phương pháp khuyến nghị lọc cộng tác được sử dụng. Kết quả phân loại bình luận được tích hợp với phân hệ khuyến nghị để tăng cường thông tin tới người dùng trước khi ra quyết định lựa chọn sản phẩm, và khắc phục vấn đề "người dùng mới". Nghiên cứu đề xuất cách xây dựng ma trận xếp hạng ảo (Virtual Rating) từ điểm phân loại bình luận thay cho xếp hạng thực (Real Rating) để khắc phục vấn đề "thưa thớt dữ liệu".

Abstract

Unstructured data generated by customers on e-commerce websites become an important matter in research and development of online recommendation system. It's assisting for online customers in making purchasing decisions. These types of data are providing new research directions for solving the challenges of traditional recommendation

* Tác giả liên hệ.
Email: lttuan@ictu.edu.vn (Lê Triệu Tuấn), hoanpm@neu.edu.vn (Phạm Minh Hoàn).
Trích dẫn bài viết: Lê Triệu Tuấn, & Phạm Minh Hoàn. (2022). Mô hình hệ thống khuyến nghị sản phẩm trong kinh doanh trực tuyến dựa vào khai thác dữ liệu phi cấu trúc. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á, 33*(12), 21–38.

Unstructured data mining.	systems. The study proposes a model of an online recommendation system based on exploiting customers' comments. The system consists of two phases, the first is the unstructured data mining process and the second phase implements product recommendations according to the collaborative filtering model approach. Comment classification results integrated with the recommendation module to enhance information to users before they are making product selection decisions and overcome the problem of the new users. Therefore, the study proposes a way to build a virtual rating from the sentiment classification score instead of the real rating to overcome the problem of sparse data.

1. Giới thiệu

Hệ thống khuyến nghị sản phẩm là một dạng hỗ trợ ra quyết định, cung cấp các giải pháp khuyến nghị mang tính cá nhân hóa, giúp khách hàng dễ dàng lựa chọn sản phẩm theo sở thích của mình (Chen và cộng sự, 2015). Hệ thống khuyến nghị sản phẩm mang lại nhiều lợi ích cho cả khách hàng và doanh nghiệp, giúp khách hàng tiết kiệm thời gian, nâng cao trải nghiệm mua sắm, giúp doanh nghiệp cải thiện doanh số, nâng cao chất lượng dịch vụ khách hàng, qua đó giữ chân khách hàng tốt hơn và nâng cao khả năng cạnh tranh (Farida, 2016; Hoàng Thị Hà & Ngô Nguyễn Thức, 2021; Hussien và cộng sự, 2021). Tuy nhiên, hiện nay vẫn đang tồn tại những thách thức chủ yếu trong hệ thống khuyến nghị làm giảm hiệu quả hoạt động, chẳng hạn như: Vấn đề "người dùng mới", vấn đề "thưa thớt dữ liệu". Những vấn đề này đã được nhiều tác giả đề xuất khắc phục nhưng chưa có một phương pháp nào có thể khắc phục hoàn toàn và vẫn đang là một hướng nghiên cứu thú vị (Sharma và cộng sự, 2022a).

Theo cách truyền thống, sở thích của khách hàng có thể được thu thập từ những bảng hỏi (Pu và cộng sự, 2012). Tuy nhiên, cách thức này thường làm tốn thời gian của khách hàng, thông tin thu được thường không đầy đủ (Asani và cộng sự, 2021). Với sự phát triển của internet và thương mại điện tử, khi tham gia mua sắm, khách hàng thường để lại một lượng lớn các bình luận dạng văn bản trên các hệ thống kinh doanh trực tuyến (He và cộng sự, 2016). Các bình luận này còn được gọi là một dạng dữ liệu phi cấu trúc (Unstructured Data). Hàm chứa trong các bình luận đó có các sở thích của người dùng. Do đó, khai thác các bình luận này sẽ biết được sở thích, xu hướng mua sắm của khách hàng. Từ đó, có thể gợi ý cho khách hàng mới những sản phẩm được nhiều người nhận xét tích cực nhất.

Việc khách hàng ra quyết định lựa chọn các sản phẩm sau khi hệ thống đưa ra danh sách sản phẩm được khuyến nghị không chỉ phụ thuộc vào sở thích của chính khách hàng đó, mà nó có thể bị ảnh hưởng bởi ý kiến của những khách hàng khác. Theo Sharma và cộng sự (2022b), trước khi ra quyết định lựa chọn sản phẩm, khách hàng thường có xu hướng tham khảo những bình luận, những xếp hạng mặt hàng đó của những khách hàng khác. Vì vậy, khai thác những dữ liệu phi cấu trúc này còn giúp khách hàng trong vấn đề ra quyết định lựa chọn sản phẩm.

Xuất phát từ những lý do trên, nhóm tác giả đề xuất mô hình hệ thống khuyến nghị sản phẩm trực tuyến dựa vào khai thác dữ liệu phi cấu trúc (Online Recommendation System Based on Untructured Data Mining – ORSUM). Mô hình hiển thị kết quả phân loại bình luận của các sản phẩm được khuyến

nghị giúp khách hàng tham khảo trước khi ra quyết định lựa chọn và khắc phục vấn đề "thưa thớt dữ liệu" khi hệ thống có số lượng người dùng xếp hạng các mặt hàng quá ít.

Nghiên cứu của nhóm tác giả đóng góp một phần vào phương pháp xây dựng hệ thống khuyến nghị sản phẩm trực tuyến kết hợp khai thác bình luận bằng tiếng Việt của khách hàng trong bối cảnh thương mại điện tử phát triển hiện nay. Ra quyết định lựa chọn mặt hàng của người dùng có thể thay đổi bởi việc hiển thị kết quả phân loại bình luận cùng với danh sách mặt hàng được khuyến nghị. Điểm phân loại được chuẩn hóa có thể thay thế những xếp hạng thực của người dùng trên các sản phẩm trong hệ thống khuyến nghị lọc cộng tác.

Cấu trúc còn lại của bài báo được tổ chức như sau: Phần 2 khảo sát những nghiên cứu liên quan; phần 3 mô tả hệ thống khuyến nghị được đề xuất; phần 4 kiểm nghiệm và đánh giá hệ thống; phần 5 kết quả và thảo luận; phần 6 kết luận và nhận định hướng nghiên cứu tiếp theo.

2. Tổng quan nghiên cứu

Hệ thống khuyến nghị tìm cách xác định sở thích của người dùng để đề xuất cho họ tập danh sách các mặt hàng phù hợp với sở thích của họ. Có một số cách để xác định sở thích của người dùng bao gồm: Xác định theo bảng hỏi (Seo và cộng sự, 2020), theo xếp hạng (Juan và cộng sự, 2019), khai thác ý kiến và phân tích tình cảm người dùng (Abbasi-Moud và cộng sự, 2021; Thái Kim Phụng và cộng sự, 2020; Ziani và cộng sự, 2017). Trong phương pháp bảng hỏi, đặc trưng sở thích của người dùng được xác định thông qua việc trả lời một số dạng câu hỏi dưới dạng bảng hỏi (Miao và cộng sự, 2016). Ví dụ, ở hệ thống khuyến nghị ẩm thực – khi truy cập vào hệ thống, người truy cập được yêu cầu chọn giá và loại thức ăn mong muốn trong danh sách các tùy chọn (Tran và cộng sự, 2018). Việc xác định theo phương pháp này có nhược điểm là sở thích của người dùng có thể không phù hợp với các câu hỏi được thiết kế sẵn. Bên cạnh đó, còn dựa trên điểm số người dùng xếp hạng cho mỗi sản phẩm, những hệ thống khuyến nghị lọc cộng tác truyền thống được phát triển (Su & Khoshgoftaar, 2009). Sở thích của người dùng được xác định bởi mức độ tương đồng với những người dùng khác trong hệ thống cùng xếp hạng cho các sản phẩm. Các giá trị xếp hạng này được biểu diễn bởi ma trận thể hiện mối quan hệ giữa người dùng (User) và sản phẩm (Item) (hay còn gọi là ma trận Utility). Thực tế, ma trận này rất thiếu những giá trị điểm số xếp hạng như vậy (Al-Bakri & Hashim, 2018). Do đó, những dạng hệ thống khuyến nghị lọc cộng tác truyền thống này thường gặp phải vấn đề "thưa thớt dữ liệu".

Sự gia tăng số lượng bình luận của người dùng trên các website thương mại điện tử đã mang lại tiềm năng lớn để xác định các đặc trưng sở thích của người dùng. Việc xác định dựa vào phân tích tình cảm có thể cung cấp thông tin linh hoạt và chính xác hơn so với cách tiếp cận trước đây. Một số hệ thống khuyến nghị dựa trên việc khai thác bình luận của khách hàng (Cambria và cộng sự, 2017; Fang & Zhan, 2015; Regina & Sengottuvelan, 2021; Sasikala & Lourdusamy, 2018; Ziani và cộng sự, 2017). Những hệ thống này xác định sở thích dựa vào phân tích tình cảm người dùng bởi các phương pháp học máy (Tuan & Hoan, 2021) và dựa trên từ điển (Phu và cộng sự, 2018) được sử dụng rộng rãi.

Leung và cộng sự (2006) đề xuất cách tiếp cận suy luận xếp hạng để tích hợp phân tích tình cảm và mô hình khuyến nghị lọc cộng tác. Leung và cộng sự (2006) thực hiện chuyển đổi sở thích của người dùng được xác định thông qua khai thác bình luận thành các thang số phù hợp với mô hình lọc

cộng tác. Liu và cộng sự (2020) tìm mối liên hệ giữa xếp hạng và các đánh giá bởi phương pháp phân tích tình cảm để bổ sung giữa xếp hạng và đánh giá giúp cải thiện vấn đề thưa thớt dữ liệu. Nghiên cứu của Sasikala và Lourdusamy (2018) đề xuất mô hình dự đoán phân loại bình luận của những người dùng không xếp hạng sản phẩm. Theo đó, mô hình thực hiện nhóm các bình luận thành hai nhóm: Một nhóm bao gồm các bình luận mà người dùng có xếp hạng sản phẩm, và nhóm còn lại bao gồm các bình luận mà người dùng không xếp hạng. Sau đó, thực hiện phân tích tình cảm cho nhóm thứ hai để dự đoán xếp hạng sản phẩm.

Việc chỉ dựa vào một tiêu chí duy nhất là các điểm số xếp hạng của khách hàng để đưa ra khuyến nghị thì không đạt được hiệu suất cao (Al-Ghuribi & Noah, 2019) bởi vì nó không phản ánh được hết sự cảm nhận mặt hàng đằng sau hành động chấm điểm đó. Cho nên, tìm hiểu thêm những cảm nhận từng khía cạnh đặc trưng của sản phẩm được khách hàng miêu tả trong các bình luận sẽ cho kết quả khuyến nghị khách quan hơn. Vấn đề ma trận Utility rất thiếu những giá trị xếp hạng của người dùng làm cho hiệu quả khuyến nghị thấp, hay thậm chí hệ thống không thể hoạt động. Và những xếp hạng này không thể hiện rõ quan điểm của khách hàng khi nhận định sản phẩm dựa trên các tiêu chí, điều này gây khó khăn trong việc giải thích các đánh giá này. Đánh giá của khách hàng ở dạng văn bản mang tính định tính và những xếp hạng được coi là thông tin định lượng. Thông tin định tính bổ sung cho thông tin định lượng vì người đánh giá phải giải thích lý do tại sao họ đánh giá sản phẩm theo cách họ làm. Do đó, bên cạnh những xếp hạng chính thức của người dùng làm đầu vào chính cho hệ thống, các nội dung bình luận của người dùng được khai thác, phân loại để bổ sung đầu vào cho hệ thống. Những bình luận có thể được khai thác, phân loại bởi phương pháp học máy (Asani và cộng sự, 2021; Lê Triệu Tuấn & Đàm Thị Phương Thảo, 2022).

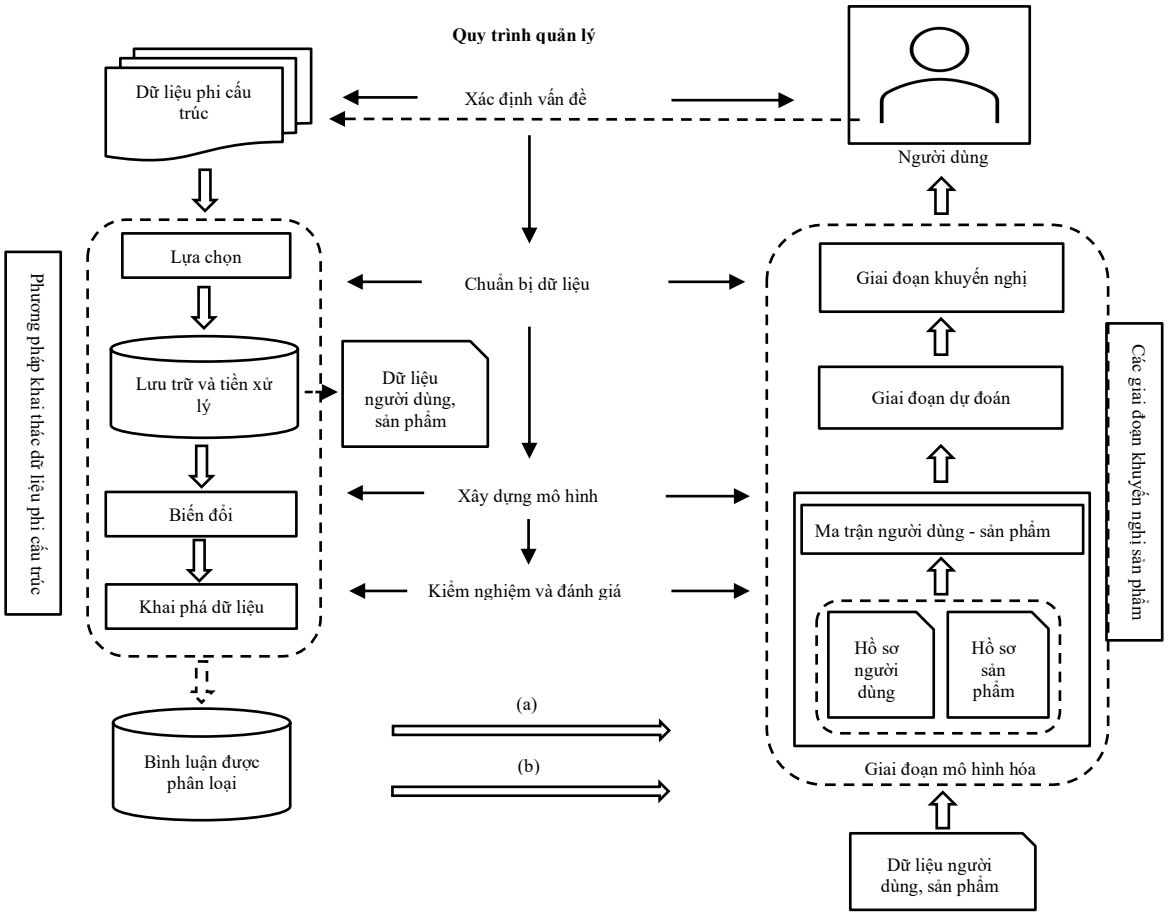
Nghiên cứu này khác với các nghiên cứu trên ở điểm nhóm tác giả thực hiện phân loại các bình luận dựa trên phân tích tình cảm, sau đó tính điểm số phân loại tương đối cho từng sản phẩm và hiển thị cùng với danh sách sản phẩm được khuyến nghị để tăng khả năng ra quyết định của người dùng. Các điểm số tương đối được nhóm tác giả chuẩn hóa về thang số phù hợp với xếp hạng thực (Real Rating) và sử dụng làm các xếp hạng ảo (Virtual Rating) trong mô hình lọc cộng tác.

3. Mô hình nghiên cứu đề xuất

Xuất phát từ cơ sở lý thuyết và các nghiên cứu liên quan, mô hình nghiên cứu tổng quát ORSUM được đề xuất như Hình 1, bao gồm hai phân hệ: Phân hệ khai thác dữ liệu phi cấu trúc và phân hệ khuyến nghị. Quy trình thực hiện theo chuẩn công nghiệp CRIP-DM (Cross Industry Standard Process for Data Mining) bao gồm các giai đoạn: Hiểu nhu cầu của doanh nghiệp (Business Understanding), hiểu về dữ liệu (Data Understanding), chuẩn bị dữ liệu (Data Preparation), xây dựng mô hình (Modeling), đánh giá mô hình (Evaluation), và triển khai giải pháp (Deployment) (Heilig và cộng sự, 2020).

Trường hợp (a): Tích hợp kết quả phân loại bình luận vào phân hệ khuyến nghị để hỗ trợ khách hàng ra quyết định lựa chọn mặt hàng và khắc phục vấn đề "người dùng mới".

Trường hợp (b): Tính điểm phân loại (Polarity Scores) cho mỗi bình luận của người dùng trên mỗi sản phẩm để có được những xếp hạng ảo khắc phục vấn đề "thưa thớt dữ liệu".



Hình 1. Mô hình nghiên cứu tổng quát

3.1. Phân hệ khai thác dữ liệu phi cấu trúc

Bước 1. Lựa chọn: Bước này thực hiện thu thập dữ liệu, nghiên cứu thực nghiệm với dữ liệu được thu thập trên trang thương mại điện tử lazada.vn. Người dùng có thể bình luận trực tiếp trên các sản phẩm hoặc thông qua các bình luận của người khác. Nghiên cứu chỉ thu thập các bình luận trực tiếp trên sản phẩm vì những bình luận gián tiếp được nhận định là ít tác động tới việc đánh giá sản phẩm của người dùng (Papadakis và cộng sự, 2017). Quá trình thu thập (Crawling Data) được thực hiện bởi thư viện Selenium (một thư viện của Python) cho phép mở trình duyệt và thao tác trên đó. Đây là phương pháp thu thập nội dung dựa vào cấu trúc Hypertext Markup Language (HTML) của các trang website (Lê Triệu Tuấn & Đàm Thị Phương Thảo, 2022). Selenium truy cập đến các phần tử của trang website thông qua các thuộc tính định danh của HTML để thu thập dữ liệu. Dữ liệu thu thập được tổ chức bao gồm các trường: Mã người dùng (UserID), mã sản phẩm (ItemID), mã phân loại (CateID), điểm số xếp hạng (Rating), ngày bình luận, và nội dung bình luận.

Bước 2. Lưu trữ và tiền xử lý: Dữ liệu sau khi thu thập sẽ được lưu trữ ở tệp có phần mở rộng là .csv (Comma Separated Values), các thông tin được phân tách nhau bởi dấu phẩy. Tiền xử lý dữ liệu được thực hiện bán tự động, loại bỏ những dữ liệu bị khuyết, những câu không có ý nghĩa, những câu

không phải tiếng Việt, những bình luận không chứa thông tin cần thiết, dấu dư thừa bởi thư viện của Python. Sau đó, thực hiện gán nhãn dữ liệu theo phương pháp của Liu và cộng sự (2017) dựa vào điểm số xếp hạng của khách hàng, kết hợp với gán nhãn thủ công. Mỗi bình luận có điểm xếp hạng từ 1 đến 5, nhóm tác giả nhận thấy rằng những bình luận có điểm xếp hạng rating >= 3 là tích cực (Positive), ngược lại rating < 3 là tiêu cực (Negative). Những bình luận chưa có điểm số xếp hạng sẽ được gán nhãn thủ công.

Bước 3. Biến đổi: Tách câu thành các từ hoặc từ ghép có nghĩa bằng thư viện Underthesea (Anh, 2018) và tính toán mức độ quan trọng của các từ bằng phương pháp TF-IDF (Arroyo-Fernández và cộng sự, 2019; Mee và cộng sự, 2021) để véc-tơ hóa văn bản. TF-IDF của từ khóa w_i trong văn bản d được tính bằng công thức sau:

$$\text{TF\_IDF} = f_{id} \times \log\frac{N}{n_i} \quad (1)$$

Trong đó,

f_{id}: Tần suất xuất hiện của từ khóa w_i trong văn bản d

$$f_{id} = \frac{\textbf{số lần từ khóa } \boldsymbol{w_i} \textbf{ xuất hiện trong văn bản } \boldsymbol{d}}{\textbf{tổng số từ trong văn bản } \boldsymbol{d}} \quad (2)$$

N: tổng số văn bản trong tập dữ liệu mẫu;

n_i: số văn bản có từ khóa w_i.

Bước 4. Khai phá dữ liệu: Bước này thực hiện phân tích tình cảm nhằm phân loại bình luận của mỗi người dùng trên mỗi sản phẩm thành hai mức cảm xúc tích cực và tiêu cực, gồm hai công đoạn: (1) Huấn luyện (Training) và thử nghiệm (Testing); và (2) dự đoán (Prediction) và tính điểm phân loại (Polarity Scrores) cho các sản phẩm. Quá trình thực hiện được mô tả như Hình 2:

Huấn luyện và thử nghiệm: Tại bước này, thực hiện theo phương pháp Hold-out (Omary & Mtenzi, 2010), dữ liệu thực nghiệm được chia theo tỷ lệ 70% dùng để huấn luyện, và 30% dùng để thử nghiệm. Các mô hình phân loại thuộc nhóm học máy có giám sát được áp dụng bao gồm: Support Vector Machine (SVM), Naïve Bayes (NB), Neural Network (NN), Logistic Regression (LR), Random Forest (RF), và Decision Tree (DT). Những mô hình này sẽ học từ dữ liệu đã được gán nhãn. Sau đó, thử nhiệm và đánh giá để lựa chọn một mô hình có độ chính xác cao nhất.

Bảng 1.

Mô tả một số mô hình học máy có giám sát được sử dụng

Mô hình	Nội dung chính	Thư viện, hàm	Nguồn tham khảo
Decision Tree	Mô hình này thực hiện phân loại văn bản bằng cách xây dựng một hệ thống dạng cây quyết định. Cây quyết định có dạng cây nhị phân, mỗi nút biểu diễn một đặc trưng của từ (Word), mỗi nhánh biểu diễn một quy luật (Rule), và mỗi lá biểu diễn một kết quả.	Scikit-learn, sklearn.svm.SVC()	Baharudin và cộng sự (2010), Salzberg (1994)
Random Forest	Mô hình này chứa nhiều cây quyết định, mỗi cây quyết định có thể sẽ khác nhau. Kết quả dự đoán được tổng hợp từ các cây quyết định này.	Scikit-learn, sklearn.ensemble.RandomForestClassifier()	Baharudin và cộng sự (2010)

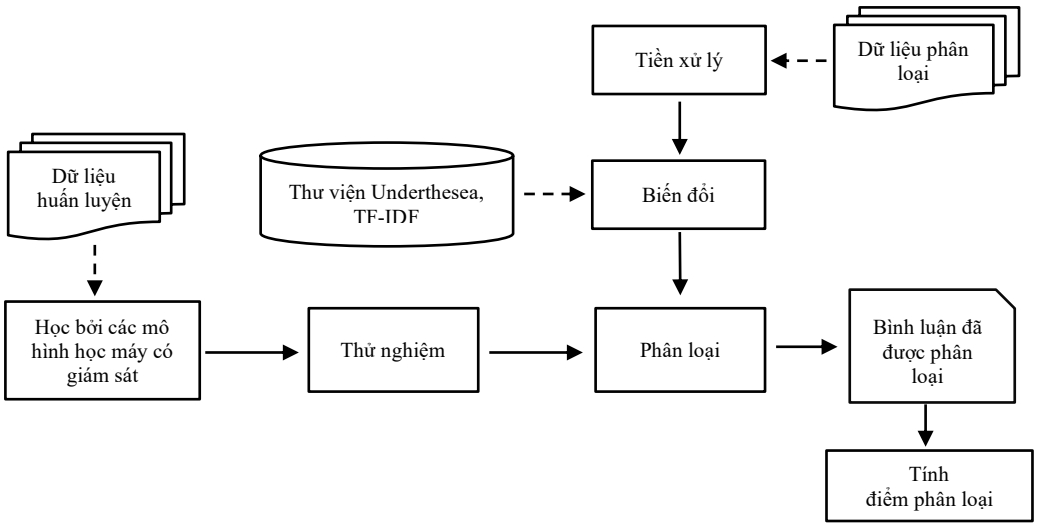
Mô hình	Nội dung chính	Thư viện, hàm	Nguồn tham khảo
Naive Bayes	Mô hình Naive Bayes phân loại các mẫu đầu vào dựa trên định lý Bayes và tính độc lập của các phần tử. Thuật toán Naive Bayes tính xác suất cho các phần tử, sau đó chọn kết quả với xác suất cao nhất.	Scikit-learn, sklearn.naive_bayes.GaussianNB()	Baharudin và cộng sự (2010), Berrar và cộng sự (2019)
Neural Network	Mô hình phân loại mạng nơ-ron bao gồm một số lượng lớn các nơ-ron được kết nối với nhau. Mô hình thiết lập một mạng nơ-ron cho mỗi văn bản trong danh mục phân loại, đầu vào đặc trưng cho véc-tơ từ, và đầu ra là danh mục văn bản được phân loại.	Scikit-learn, sklearn.neural_network.MLPClassifer()	Baharudin và cộng sự (2010)
Support Vector Machine	Mô hình SVM thực hiện phân loại dữ liệu thành các lớp phân tách nhanh bởi một siêu mặt phẳng. Trong nghiên cứu này, dữ liệu được chia thành hai lớp Positive và Negative, lúc này siêu mặt phẳng được thuật toán tìm thấy sẽ phân tách hai lớp này.	Scikit-learn, sklearn.svm.SVC()	Baharudin và cộng sự (2010), Lê Thị Minh Nguyện (2019), Graff và cộng sự (2022)
Logistic Regression	Hồi quy logistic là một phương pháp phân tích thống kê, dự đoán giá trị dữ liệu dựa trên các quan sát trước đó của tập dữ liệu. Mục đích của hồi quy logistic là ước tính xác suất của các sự kiện, từ đó dự đoán xác suất của các kết quả. Dữ liệu văn bản đầu vào của thuật toán đã được gán nhãn (Positive và Negative). Đầu ra là xác suất dữ liệu rơi vào nhãn tương ứng.	Scikit-learn, sklearn.linear_model.LogisticRegression()	Ambris và cộng sự (2022)

Tính điểm phân loại: Dữ liệu bình luận trên mỗi sản phẩm cần phân loại cũng sẽ được tiền xử lý sau đó thực hiện biến đổi và đưa vào mô hình phân loại đã được lựa chọn.

Nhóm tác giả thực hiện tính điểm phân loại (R_{score}) trên mỗi sản phẩm *i* để làm giá trị đầu vào cho phân hệ khuyến nghị. Thông thường, trên hệ thống website sẽ cho phép người dùng có thể bình luận nhiều lần với mỗi mặt hàng, nhưng sẽ được gộp lại và tính là một bình luận. R_{score} được suy ra từ các phân loại tích cực và tiêu cực của mỗi đoạn văn bản bởi công thức sau:

$$R_{score} = \frac{\boldsymbol{P_{score} - N_{score}}}{\boldsymbol{P_{score} + N_{score}}} \quad (3)$$

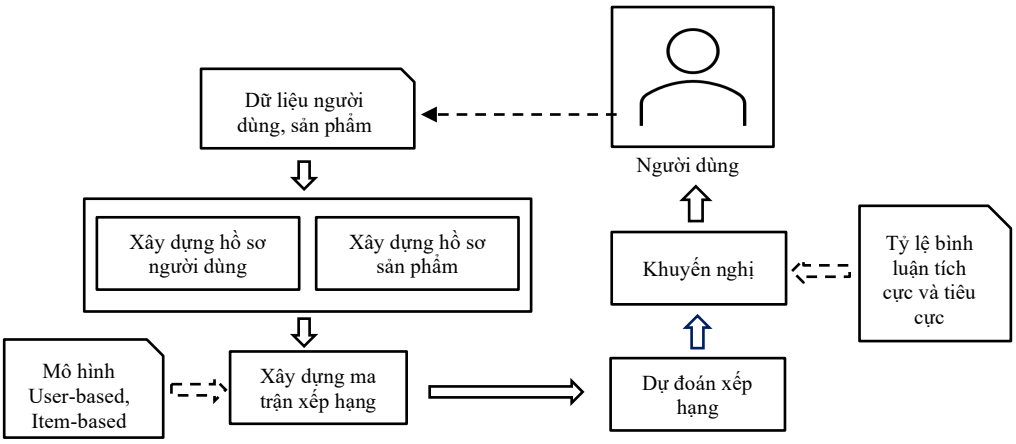
Trong đó, P_{score} và N_{score} là số bình luận tích cực và tiêu cực trên *i*. R_{score} nằm trong đoạn [–1, 1] (–1 thể hiện đoạn bình luận R hoàn toàn tiêu cực, 1 hoàn toàn tích cực).



Hình 2. Quy trình huấn luyện và phân loại bình luận

3.2. Phân hệ khuyến nghị

Dạng hệ thống được sử dụng trong nghiên cứu này là hệ thống khuyến nghị lọc cộng tác (Collaborative Filtering), dựa theo mô hình thuật toán K láng giềng (Neighborhood-Based hay còn gọi là Memory-Based). Nghiên cứu tiến hành thực nghiệm theo hai cách tiếp cận, dựa trên mối quan hệ tương đồng có cùng sở thích giữa người dùng hiện tại với những người dùng khác trong hệ thống (User-Based) và dựa trên độ tương đồng giữa sản phẩm được đề xuất khuyến nghị với những sản phẩm khác trong hệ thống (Item-Based). Mối quan hệ được thể hiện bởi điểm số Rating xếp hạng sản phẩm của người dùng, điểm số Rating này có giá trị từ 1 đến 5.



Hình 3. Phân hệ khuyến nghị của ORSUM tích hợp hiển thị tỷ lệ phân loại bình luận

- *Xây dựng hồ sơ người dùng, sản phẩm*

Hồ sơ người dùng gồm các đặc trưng: Mã người dùng, những mặt hàng người dùng đã xếp hạng. Hồ sơ sản phẩm bao gồm các đặc trưng: Mã sản phẩm, mã người dùng đã xếp hạng, điểm số xếp hạng.

- *Xây dựng ma trận Utility*

Mỗi dòng của ma trận Utility là một véc tơ chứa các phần tử bao gồm các giá trị điểm số xếp hạng tương ứng của các sản phẩm. Sau khi ma trận Utility được tạo lập, hệ thống tính độ tương đồng giữa các Users với mô hình User-Based hoặc giữa các Items với mô hình Item-Based.

Cách tiếp cận theo mô hình User-Based xác định độ tương đồng giữa hai người dùng thông qua việc so sánh các đánh giá của họ trên cùng sản phẩm, sau đó dự đoán xếp hạng trên sản phẩm *i* bởi người dùng *u* thông qua các xếp hạng của những người dùng tương tự *u'*. Độ tương đồng (Sim) giữa người dùng *u* và *u'* được tính theo công thức Cosine (Herlocker và cộng sự, 2002).

$$\text{Sim (u, u')} = \frac{\sum_{i \in I_{uu'}} r_{ui} \times r_{u'i}}{\sqrt{\sum_{i \in I_{uu'}} r_{ui}^2} \sqrt{\sum_{i \in I_{uu'}} r_{u'i}^2}} \quad (4)$$

Trong đó,

r_{ui} và $r_{u'i}$: Xếp hạng của người dùng *u* và *u'* trên sản phẩm *i* tương ứng;

$I_{uu'}$: Tập các sản phẩm được đánh giá xếp hạng bởi cả người dùng *u* và người dùng *u'*.

Tương tự cho cách tiếp cận theo mô hình Item-Based, độ tương đồng giữa hai sản phẩm *i* và *i'* được xác định như sau:

$$\text{Sim (i, i')} = \frac{\sum_{u \in U_{ii'}} r_{ui} \times r_{ui'}}{\sqrt{\sum_{u \in U_{ii'}} r_{ui}^2} \sqrt{\sum_{u \in U_{ii'}} r_{ui'}^2}} \quad (5)$$

Trong đó, $U_{ii'}$ Tập các người dùng có đánh giá xếp hạng trên cả hai sản phẩm *i* và *i'*.

- *Dự đoán xếp hạng*

Sau khi tính toán độ tương đồng giữa các người dùng, dự đoán xếp hạng ($\hat{r}_{ui}$) của người dùng *u* lên sản phẩm *i* được xác định theo công thức (Resnick và cộng sự, 1994):

Với mô hình User-Based:

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{u' \in K_u} sim(u,u') \times (r_{u'i} - \bar{r}_{u'})}{\sum_{u' \in K_u} |sim(u,u')|} \quad (6)$$

Trong đó,

$\bar{r}_u$ và $\overline{r_{u'}}$: Giá trị đánh giá xếp hạng trung bình trên tất cả các sản phẩm của người dùng *u* và *u';*

K_u: Số người dùng lân cận *u* (*k* láng giềng của *u*).

Với mô hình Item-Based:

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_{i' \in K_i} sim(i,i') \times (r_{u'i} - \bar{r}_{i'})}{\sum_{i' \in K_i} |sim(i,i')|} \quad (7)$$

Trong đó,

$\bar{r}_i$ và $\overline{r_{i'}}$: Giá trị đánh giá xếp hạng trung bình của tất cả người dùng trên sản phẩm *i* và *i';*

K_i: Số sản phẩm lân cận *i* (*k* láng giềng của *i*).

- *Khuyến nghị sản phẩm*

Hệ thống tìm tất cả những sản phẩm có độ tương tự cao nhất với sản phẩm người dùng hiện tại để khuyến nghị. Nhóm tác giả nhận thấy rằng, không phải lúc nào khách hàng cũng quan tâm tới các sản phẩm được khuyến nghị và đưa ra lựa chọn khi được khuyến nghị. Do đó, để tăng thêm giá trị cho những khuyến nghị và thu hút thêm sự chú ý của khách hàng, hệ thống cung cấp thêm thông tin về tỷ lệ số bình luận tích cực trên mỗi sản phẩm được khuyến nghị để khách hàng tham khảo ra quyết định.

Trong trường hợp dữ liệu của ma trận Utility rất thưa làm cho hệ thống không thể dự đoán xếp hạng sản phẩm. Lúc này, nhóm tác giả đề xuất các giá trị xếp hạng thực của ma trận Utility được thay thế bởi các giá trị R_{score} và gọi là các xếp hạng ảo. Việc xác định độ phù hợp của các xếp hạng ảo này sẽ được đánh giá bởi kỳ vọng (Expectation) với các xếp hạng thực có trong bộ dữ liệu và phương sai (Variance) thể hiện sự khác biệt giữa xếp hạng thực và kỳ vọng.

Mỗi người dùng *u* có một vector ảo, ký hiệu $Vr_{ui}(u)$. Mỗi Vr_{ui} có giá trị $R_{score}(N)$ được chuẩn hóa từ R_{score} về thang điểm xếp hạng thực của từng hệ thống bán hàng (R_{rating}). Trong nghiên cứu này, nhóm tác giả đang thực nghiệm là thang điểm 5.

$$R_{score}(N) = \left|\frac{\mathbf{R_{score}}}{\mathbf{Max(R_{score})}}\right| \times \mathbf{R_{rating}} \quad (8)$$

Nghiên cứu không xét trường hợp $R_{score} = 0$ vì trường hợp này người dùng *u* không quan tâm tới sản phẩm *i* (không bình luận) và nó không có ý nghĩa để đưa ra tham khảo khuyến nghị.

4. Kiểm nghiệm và đánh giá

4.1. Dữ liệu thực nghiệm

Nghiên cứu đã thu thập được 32.148 bình luận trên 1.407 sản phẩm thuộc 12 nhóm mặt hàng. Để có được những xu hướng mới của khách hàng, nhóm tác giả thực hiện lọc và chỉ giữ lại dữ liệu được bình luận trong khoảng thời gian năm năm gần nhất (từ năm 2017 đến năm 2022). Sau khi tiền xử lý, nhóm tác giả thu được 31.728 bình luận, thực hiện gán nhãn bình luận tích cực và bình luận tiêu cực. Các bình luận trung tính (Neutral) sẽ không được sử dụng vì không có ý nghĩa trong việc hỗ trợ khuyến nghị.

4.2. Đánh giá và lựa chọn mô hình phân loại bình luận

Độ chính xác của các mô hình phân loại bình luận được đánh giá bằng ma trận nhầm lẫn (Confused Matrix) thông dụng (Kulkarni và cộng sự, 2020), bao gồm bốn yếu tố để xác định các chỉ số đo lường hiệu suất mô hình như sau:

TP (True Positive): Số lượng bình luận của lớp tích cực được dự đoán đúng là tích cực.

FP (False Positive): Số lượng bình luận của lớp tiêu cực bị dự đoán nhầm thành tích cực.

TN (True Negative): Số lượng bình luận của lớp tiêu cực được dự đoán đúng là tiêu cực.

FN (False Negative): Số lượng bình luận của lớp tích cực bị dự đoán nhầm thành tiêu cực.

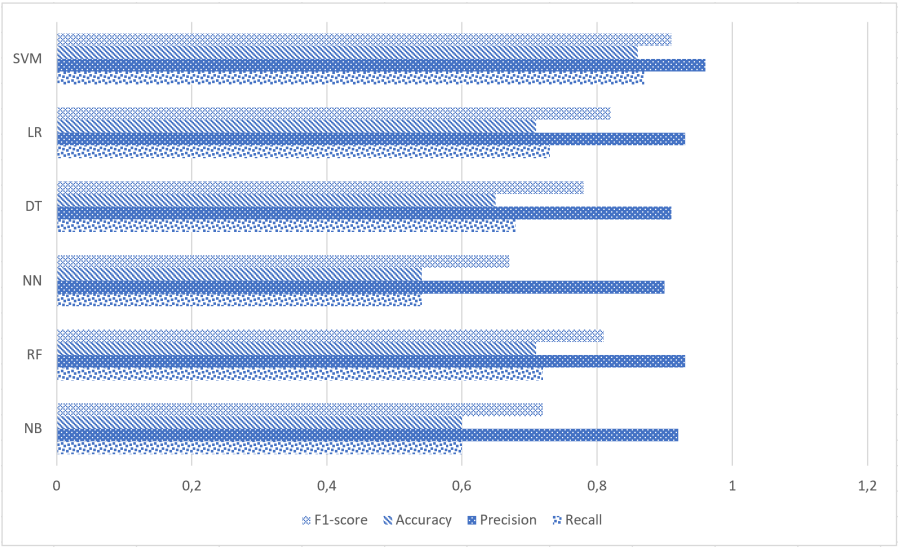
Bốn chỉ số đó là: Độ chính xác (Accuracy), độ hội tụ (Precision), độ bao phủ (Recall), và giá trị trung bình điều hòa ($F1_{score}$). $F1_{score}$ có giá trị càng cao thì mô hình phân loại càng chính xác.

$$\text{Accuracy} = \frac{(\text{TP}+ \text{TN})}{(\text{TP}+ \text{FP}+ \text{FN}+ \text{TN})} \quad (9)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+ \text{FP}} \quad (10)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+ \text{FN}} \quad (11)$$

$$\text{F1}_{\text{score}} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+ \text{Recall}} \quad (12)$$



Hình 4. Kết quả các chỉ số của các mô hình phân loại

Hình 4 cho thấy rằng mô hình SVM có *F1-score* lớn nhất so với các mô hình còn lại (91%). Nhóm tác giả sử dụng mô hình SVM để thực hiện xây dựng hệ thống khuyến nghị sản phẩm trực tuyến dựa vào phân tích dữ liệu phi cấu trúc.

4.3. Đánh giá độ tin cậy của mô hình

Nhóm tác giả sử dụng Root Mean Squared Error (RMSE) để xác định độ lỗi của mô hình dự đoán khuyến nghị (Papadakis và cộng sự, 2017; Trần Nguyễn Minh Thư & Phạm Xuân Hiền, 2016; Nguyễn Thái Nghe, 2016), RMSE được xác định bởi công thức sau:

$$\text{RMSE} = \sqrt{\frac{1}{|D^{test}|}\sum_{u,i\,\epsilon\,D^{test}}(r_{ui} - \hat{r}_{ui})^2} \quad (13)$$

Trong đó,

D^{test}: Tập dữ liệu kiểm tra;

$|D^{test}|$: Số quan sát trong tập D^{test} ($|D^{test}| = 9.518$);

r_{ui}: Dự đoán xếp hạng của người dùng u trên tập D^{test};

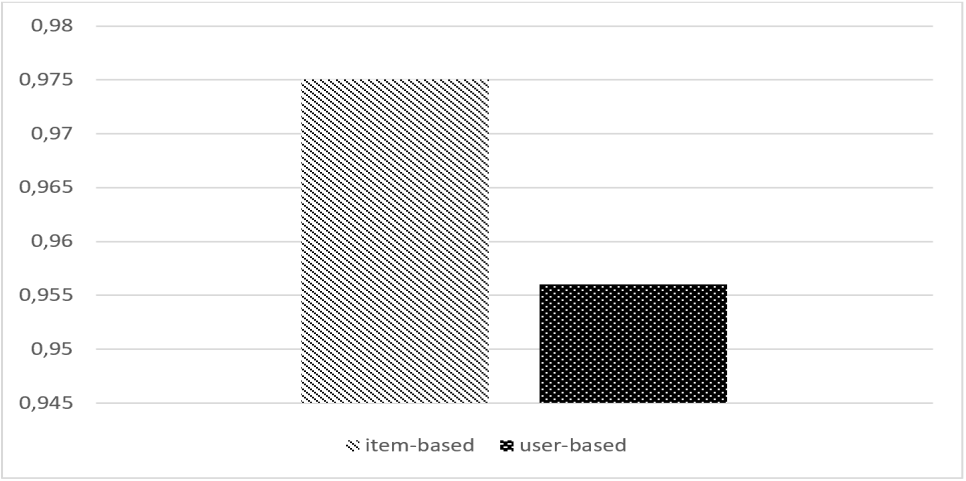
$\hat{r}_{ui}$: Dự đoán xếp hạng của người dùng u trên sản phẩm i xác định theo công thức (6) và (7).

Giá trị RMSE càng nhỏ tức là sai số càng bé thì tính chính xác của mô hình dự đoán càng tin cậy.

Bảng 2.
Một số giá trị dự đoán xếp hạng của người dùng

Mô hình Item-based			Mô hình User-based		
r_{ui}	$\hat{r}_{ui}$	$(r_{ui} - \hat{r}_{ui})^2$	r_{ui}	$\hat{r}_{ui}$	$(r_{ui} - \hat{r}_{ui})^2$
3	3,391	0,152	3	3,192	0,037
4	3,763	0,058	4	3,791	0,044
5	4,353	0,423	5	4,441	0,312
4	3,372	0,397	4	3,022	0,956
3	3,794	0,624	3	3,663	0,440
5	3,872	1,277	5	4,081	0,845
4	3,723	0,078	4	3,934	0,004
4	3,582	0,176	4	2,923	1,160
5	4,041	0,922	5	4,082	0,843
5	3,450	2,403	5	3,281	2,955
4	3,783	0,048	4	3,742	0,067
4	3,732	0,073	4	3,251	0,561
4	3,252	0,563	4	3,001	0,998
4	3,761	0,058	4	3,292	0,501
4	3,864	0,020	4	3,770	0,053
3	3,594	0,348	3	3,393	0,154
3	3,952	0,903	3	3,110	0,012
5	4,144	0,740	5	4,553	0,200
1	4,262	10,628	1	4,091	9,554
5	4,051	0,903	5	4,173	0,684
3	2,480	0,270	3	2,662	0,114
2	2,873	0,757	2	2,262	0,069
4	3,593	0,168	4	4,163	0,027
2	2,662	0,436	2	2,614	0,377
2	2,530	0,250	2	2,443	0,196
1	2,841	3,386	1	2,111	1,234
3	2,843	0,026	3	3,053	0,003
4	2,514	2,220	4	4,140	0,020

Từ các giá trị định lượng xếp hạng của người dùng trong tập dữ liệu thử nghiệm và dự đoán xếp hạng trên từng sản phẩm, nhóm tác giả xác định được độ lỗi của các mô hình thuật toán dự đoán khuyến nghị như sau:



Hình 5. Độ lỗi của các mô hình dự đoán khuyến nghị

Hình 5 cho thấy rằng, mức độ sai số của mô hình User-Based nhỏ hơn nhiều so với mức độ sai số của mô hình Item-Based.

5. Kết quả và thảo luận

5.1. Khuyến nghị

Hệ thống kết hợp hiển thị danh sách 5 sản phẩm được khuyến nghị theo mô hình User-Based với tỷ lệ phân loại bình luận tích cực của những sản phẩm đó tới người dùng hiện tại.

Bảng 3.

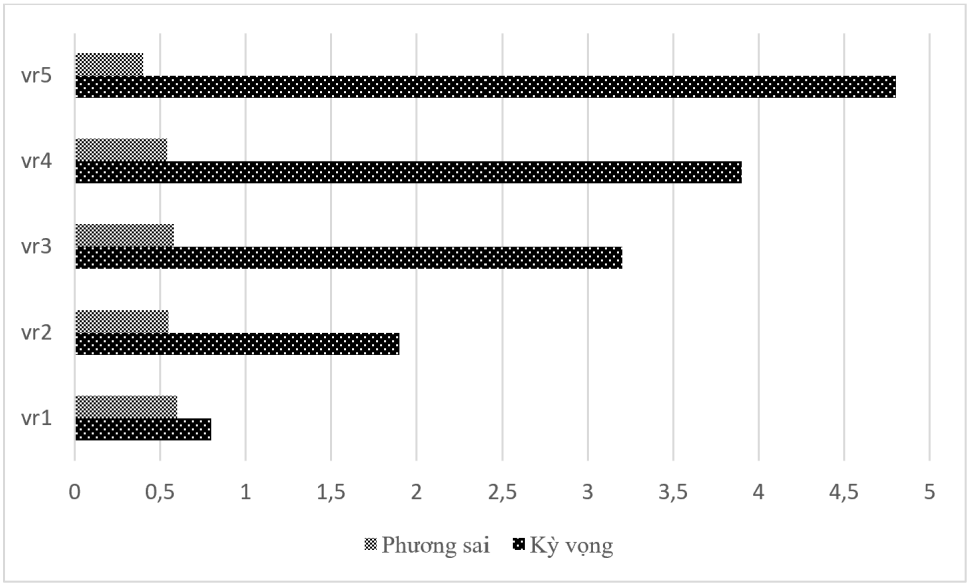
Danh sách 5 sản phẩm được khuyến nghị tới người dùng hiện tại kết hợp với tỷ lệ phân loại bình luận

STT	Mã sản phẩm	Độ tương đồng	Số bình luận tích cực	Tổng số bình luận	Tỷ lệ %
1	2__1372140635	34,15	743	926	80,24
2	2__1524844029	26,34	1.828	2.039	89,65
3	2__1256795401	14,45	1.139	1.186	96,04
4	2__1524477237	11,37	108	168	64,29
5	2__1525172201	7,98	332	431	77,03

Kết quả trong Bảng 2, mặt hàng có mã 2__1256795401 ở vị trí khuyến nghị thứ ba nhưng lại có tỷ lệ bình luận tích cực cao nhất. Điều này thể hiện rằng, mặt hàng này tuy không gần sở thích với người dùng hiện tại bằng hai mặt hàng xếp trên, nhưng được nhiều người dùng trong cộng đồng đánh giá cao. Đó cũng là một tiêu chí cho người dùng tham khảo để lựa chọn, phù hợp với xu hướng khách hàng thường đọc bình luận của những người dùng trước để ra quyết định mua hàng.

5.2. Xếp hạng ảo

Xếp hạng thực của người dùng trên một sản phẩm là một số nguyên từ 1 đến 5, 1 có nghĩa là rất không thích, còn 5 là rất thích mặt hàng đó. Sau khi thực hiện Phân hệ 1, nhóm tác giả có được các giá trị điểm số phân loại bình luận. Thực hiện chuẩn hóa các giá trị này bởi công thức (8), nhóm tác giả nhận được các xếp hạng ảo có giá trị nằm trong đoạn [0, 5]. Để xem xét mối quan hệ giữa xếp hạng ảo và xếp hạng thực, nhóm tác giả xây dựng một tập hợp các xếp hạng ảo vr_i tương ứng với mỗi danh mục điểm của xếp hạng thực và thu được 5 danh mục. Sau đó, nhóm tác giả tính kỳ vọng và phương sai cho các danh mục xếp hạng ảo này.



Hình 6. Kỳ vọng và phương sai của xếp hạng ảo đối với xếp hạng thực

Kỳ vọng các xếp hạng ảo tiệm cận gần với các xếp hạng thực và các phương sai rất nhỏ. Như vậy, có thể dùng các xếp hạng ảo này để thực hiện khuyến nghị sản phẩm như các xếp hạng thực của khách hàng. Cách này có thể khắc phục được vấn đề "thưa thớt dữ liệu".

5.3. Thảo luận kết quả

Nghiên cứu này mới chỉ thu thập dữ liệu ở dạng tĩnh, các dữ liệu sinh ra có thể được thu thập và xử lý ngay theo ngữ cảnh ở những hệ thống tương tự trong tương lai. Hỗ trợ khuyến nghị sản phẩm bởi kết quả phân loại bình luận cung cấp thêm cho người dùng một cái nhìn khách quan về sản phẩm mình đang quan tâm ở khía cạnh những người dùng trước. Sản phẩm có thể được đa số người dùng đánh giá điểm cao nhưng nó không phản ánh hết vấn đề liên quan. Những cảm nhận sẽ được bộc lộ trong những bình luận của khách hàng sau khi trải nghiệm. Do đó, một sản phẩm được xếp hạng điểm cao chưa hẳn đã tốt. Và việc tổng hợp, phân tích những bình luận sẽ củng cố thêm thông tin cho người dùng ra quyết định lựa chọn sản phẩm. Đồng thời, nó cũng có thể dùng để gợi ý sản phẩm cho những người dùng mới trong hệ thống. Việc tính toán điểm phân loại bình luận và chuẩn hóa điểm phân loại về thang đo của xếp hạng thực để xây dựng những xếp hạng ảo có kỳ vọng cao với những xếp hạng thực. Điều đó cho thấy, những phản ánh về sản phẩm ẩn chứa trong bình luận của khách hàng phần lớn có ý nghĩa tương tự như việc họ xếp hạng.

6. Kết luận và hướng nghiên cứu tiếp theo

Nhóm tác giả đã phát triển một mô hình hệ thống khuyến nghị sản phẩm trực tuyến dựa vào khai thác dữ liệu phi cấu trúc là các bình luận của khách hàng dạng văn bản tiếng Việt. Mô hình ORSUM gồm hai phân hệ, phân hệ thứ nhất bằng việc phát triển chương trình khai thác dữ liệu tự động và phân loại văn bản bằng học máy có giám sát, nhóm tác giả đã phân loại bình luận của khách hàng trên các sản phẩm thành những bình luận tích cực và bình luận tiêu cực.

Thông qua việc cung cấp các kết quả phân loại bình luận tới khách hàng được khuyến nghị, ORSUM đã cung cấp thêm thông tin cho khách hàng tham khảo trước khi ra quyết định. Các kết quả khuyến nghị này cũng có thể giúp khắc phục vấn đề "người dùng mới" bằng cách hiển thị các mặt hàng được người dùng trước đó xếp hạng cao nhất tới người dùng mới.

Qua việc tính điểm phân loại bình luận và chuẩn hóa về thang điểm xếp hạng thực, nhóm tác giả có được những xếp hạng ảo có kỳ vọng gần với xếp hạng thực. Điều đó cho thấy có thể sử dụng những xếp hạng ảo này để dự đoán khuyến nghị thay cho những xếp hạng thực khi hệ thống xuất hiện vấn đề "thưa thớt dữ liệu". Và qua đó, cũng cho nhóm tác giả thấy được sự tác động đáng kể của những bình luận tới việc đánh giá mặt hàng của khách hàng.

Ngoài lĩnh vực kinh doanh trực tuyến, nghiên cứu cũng có thể được áp dụng trong các lĩnh vực khác như: Giáo dục, y tế, truyền thông... khi các lĩnh vực này có những bình luận, đánh giá của người dùng. Cùng với mục tiêu hỗ trợ khách hàng, nghiên cứu cũng giúp nhà quản lý hiểu sâu sắc hơn xu hướng của khách hàng và nâng cao hiệu quả quản trị kinh doanh. Việc chỉ phân loại bình luận thành tích cực và tiêu cực không phản ánh hết những ẩn ý của khách hàng trong đó. Do đó, tương tương lai, nhóm tác giả sẽ nghiên cứu các phương pháp mới để tăng cường chất lượng dữ liệu cho phân loại văn bản, hiểu sâu sắc hơn những ẩn ý của khách hàng trong các bình luận để có thêm những khuyến nghị chất lượng hơn. Đặc biệt, nhóm tác giả sẽ tiếp tục nghiên cứu thêm việc khắc phục vấn đề "khởi động nguội" và "thưa thớt dữ liệu" từ phân tích dữ liệu phi cấu trúc.

Chú thích

Bài báo được phát triển nội dung từ nghiên cứu "Model of Product Recommendation System in Online Business based on Unstructured Data Mining" của nhóm tác giả "Le Trieu Tuan và Pham Minh Hoan" đã được trình bày tại Hội thảo Khoa học Quốc tế về Kinh tế và Kinh doanh Châu Á lần thứ 4 (The 4th Asia Conference on Business and Economic Studies – ACBES 2022) do Trường Đại học Kinh tế TP. Hồ Chí Minh tổ chức. Nhóm tác giả xin chân thành cảm ơn những quan tâm và góp ý cho nghiên cứu tại Hội thảo. Nhóm tác giả cam kết chỉ đăng bản tiếng Việt này trên Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á và không sử dụng bản dịch sang tiếng Anh đăng tạp chí khác.

Bài báo này được trích một phần từ kết quả nghiên cứu trong luận án tiến sĩ của Nghiên cứu sinh Lê Triệu Tuấn do giảng viên TS. Phạm Minh Hoàn hướng dẫn tại Trường Đại học Kinh tế Quốc dân.

Tài liệu tham khảo

Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Systems with Applications, 167*(2–3), 114324. doi: 10.1016/j.eswa.2020.114324

Al-Ghuribi, S. M., & Noah, S. A. M. (2019). Multi-criteria review-based recommender system–the state of the art. *IEEE Access, 7*, 169446-169468. doi: 10.1109/ACCESS.2019.2954861

Al-Bakri, N. F., & Hashim, S. H. (2018). Reducing data sparsity in recommender systems. *Al-Nahrain Journal of Science, 21*(2), 138–147.

Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj and Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings, 3*(1), 127–130.

Anh, V. (2018). Underthesea Document. In *Underthesea*. Retrieved from https://underthesea.readthedocs.io.

Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., & Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF–IDF. *Computer Speech & Language, 56*, 107–129. doi: 10.1016/j.csl.2019.01.005

Asani, E., Vahdat-Nejad, H., & Sadri, J. (2021). Restaurant recommender system based on sentiment analysis. *Machine Learning with Applications, 6,* 100114. doi: 10.1016/j.mlwa.2021.100114

Baharudin, B., Lee, L. H., Khan, K., & Khan, A. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology, 1*(1), 4–20.

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, *108*(1), 97–126.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems, 32*, 74–80. doi: 10.1109/MIS.2017.4531228

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (1999). Crisp-dm 1.0.: Step-by-step data mining guide. *CRISP-DM consortium*. Retrieved from http://www.crisp-dm.org

Chen, L., Chen, G., & Wang, F. (2015). Recommender systems based on user reviews: The state of the art. *User Modeling and User-Adapted Interaction, 25*(2), 99–154. doi: 10.1007/s11257-015-9155-5

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data, 2*(1), 5–19. doi: 10.1186/s40537-015-0015-2

Farida, K. (2016). A survey of e-commerce recommender systems. *European Scientific Journal, 12*(34), 76–89. doi: 10.19044/esj.2016.v12n34p75

Graff, M., Moctezuma, D., Miranda-Jiménez, S., & Tellez, E. S. (2022). A Python library for exploratory data analysis on twitter data based on tokens and aggregated origin–destination information. *Computers & Geosciences*, *159*, 105012. doi: 10.1016/j.cageo.2021.105012

Hoàng Thị Hà, & Ngô Nguyễn Thức. (2021). Một số phương pháp gợi ý và ứng dụng trong thương mại điện tử. *Tạp chí khoa học nông nghiệp Việt Nam, 19*(4), 520–534.

He, P., Zhu, J., He, S., Li, J., & Lyu, M. R. (2016). *An evaluation study on log parsing and its use in log mining.* Paper presented at the 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN).

Heilig, L., Stahlbock, R., & Voß, S. (2020). From digitalization to data-driven decision making in container terminals. In J. W. Böse (Ed.), *Handbook of Terminal Planning* (pp. 125–154). Cham: Springer International Publishing.

Herlocker, J., Konstan, J., & Riedl, J. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information Retrieval, 5*, 287–310. doi: 10.1023/A:1020443909834

Hussien, F. T. A., Rahma, A. M. S., & Wahab, H. B. A. (2021). *Recommendation Systems For E-commerce Systems An Overview.* Paper presented at the Journal of Physics: Conference Series.

Juan, W., Yue-xin, L., & Chun-ying, W. (2019). Survey of recommendation based on collaborative filtering. *Journal of Physics: Conference Series, 1314*(1), 012078. doi: 10.1088/1742-6596/1314/1/012078

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). 5 - Foundations of data imbalance and solutions for a data democracy. In F. A. Batarseh & R. Yang (Eds.), *Data Democracy* (pp. 83–106). Academic Press.

Leung, C. W., Chan, S. C., & Chung, F.-l. (2006). *Integrating collaborative filtering and sentiment analysis: A rating inference approach.* Paper presented at the Proceedings of the ECAI 2006 workshop on recommender systems.

Lê Thị Minh Nguyện. (2019). Text classification based on support vector machine. *Tạp chí Khoa học Đại học Đà Lạt*, *9*(2), 3–19.

Liu, H., Wang, Y., Peng, Q., Wu, F., Gan, L., Pan, L., & Jiao, P. J. N. (2020). Hybrid neural recommendation with joint deep representation learning of ratings and reviews. *Neurocomputing, 374*, 77–85.

Liu, Z., Luo, C., & Liu, G. (2017). Study on inductive effect of product information label on technology: An empirical based on the energy efficiency labeling system in China. *Science & Technology Progress and Policy*, *34*(24), 18–24.

Mee, A., Homapour, E., Chiclana, F., & Engel, O. (2021). Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit. *Knowledge-Based Systems, 228*(4), 107238. doi: 10.1016/j.knosys.2021.107238

Miao, X., Gao, Y., Chen, G., Cui, H., Guo, C., & Pan, W. (2016). SI2P: A restaurant recommendation system using preference queries over incomplete information. *Proceedings of the VLDB Endowment, 9*(13), 1509–1512. doi: 10.14778/3007263.3007296

Nguyễn Thái Nghe. (2016). *Hệ thống gợi ý: Kỹ thuật và ứng dụng* (trang 21–47). Cần Thơ: NXB Đại học Cần Thơ.

Omary, Z., & Mtenzi, F. (2010). Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. *International Journal for Infonomics*, *3*(3), 314–325.

Papadakis, H., Panagiotakis, C., & Fragopoulou, P. (2017). SCoR: A synthetic coordinate based recommender system. *Expert Systems with Applications, 79,* 8–19. doi: 10.1016/j.eswa.2017.02.025

Phu, V. N., Chau, V. T. N., Tran, V. T. N., & Dat, N. D. (2018). A Vietnamese adjective emotion dictionary based on exploitation of Vietnamese language characteristics. *Artificial Intelligence Review, 50*(1), 93–159. doi: 10.1007/s10462-017-9538-6

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

Tran, T. N. T., Atas, M., Felfernig, A., & Stettinger, M. (2018). An overview of recommender systems in the healthy food domain. *Journal of Intelligent Information Systems*, *50*(3), 501–526. doi: 10.1007/s10844-017-0469-0

Pu, P., Chen, L., Hu, R. (2012). Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Modeling and User-Adapted Interaction*, *22*(4–5), 317–355.

Regina, I. A., & Sengottuvelan, P. (2021). *Analysis of sentiments in movie reviews using supervised machine learning technique.* Paper presented at the 2021 4th International Conference on Computing and Communications Technologies (ICCCT).

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). *Grouplens: An open architecture for collaborative filtering of netnews.* Paper presented at the Proceedings of the 1994 ACM conference on Computer supported cooperative work.

Salzberg, S. L. (1994). *Programs for Machine Learning.* Kluwer Academic Publishers.

Sasikala, P., & Lourdusamy, M. I. S. (2018). Sentiment analysis and prediction of online reviews with empty ratings. *International Journal of Applied Engineering Research, 13*, 11525–11531.

Seo, Y.-D., Lee, E., & Kim, Y.-G. (2020). Video on demand recommender system for internet protocol television service based on explicit information fusion. *Expert Systems with Applications, 143*, 113045. doi: 10.1016/j.eswa.2019.113045

Sharma, D. K., Lohana, S., Arora, S., Dixit, A., Tiwari, M., & Tiwari, T. (2022a). E-commerce product comparison portal for classification of customer data based on data mining. *Materials Today: Proceedings, 51*, 166–171. doi: 10.1016/j.matpr.2021.05.068

Sharma, M., Mittal, R., Bharati, A., Saxena, D., & Singh, A. (2022b). A survey and classification on recommendation systems. *Conference: 2nd International Conference on Big Data, Machine Learning and Applications*.

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *2009*(2), 2–21.

Trần Nguyễn Minh Thư, & Phạm Xuân Hiền. (2016). Các phương pháp đánh giá hệ thống gợi ý. *Tạp chí khoa học Trường Đại học Cần Thơ, 42*, 18–27.

Tuan, L. T., & Hoan, P. M. (2021). *Method of content classification based on supervised machine learning in online comment mining of customer.* Paper presented at the Socio-Economic and Environmental Issues in Development, pp. 823–832, National Economics University.

Lê Triệu Tuấn, & Đàm Thị Phương Thảo. (2022). Phương pháp phân loại dữ liệu bình luận của khách hàng trực tuyến Việt Nam dựa vào học máy có giám sát. *Khoa học & Công nghệ, 58*(1), 49–52.

Ziani, A., Azizi, N., Schwab, D., Aldwairi, M., Chekkai, N., Zenakhra, D., & Cheriguene, S. (2017). *Recommender System Through Sentiment Analysis*. Paper presented at 2nd International Conference on Automatic Control, Telecommunications and Signals (ICATS), Algeria.