MÔ HÌNH HỆ THỐNG KHAI THÁC DỮ LIỆU PHI CẤU TRÚC HỖ TRỢ KHÁCH HÀNG RA QUYẾT ĐỊNH MUA HÀNG TRỰC TUYẾN

AN UNSTRUCTURED DATA MINING SYSTEM MODEL TO SUPPORT CUSTOMERS IN MAKING ONLINE PURCHASING DECISIONS

Lê Triệu Tuấn[1]*, Phạm Minh Hoàn[2]

[1]Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên
[2]Trường Đại học Kinh tế Quốc dân

*Tác giả liên hệ: lttuan@ictu.edu.vn

(Nhận bài: 19/9/2022; Chấp nhận đăng: 20/11/2022)

Tóm tắt - Những dạng dữ liệu phi cấu trúc được khách hàng để lại trên không gian mạng hiện nay ngày càng trở nên quan trọng đối với các doanh nghiệp kinh doanh trực tuyến. Dữ liệu đó là những bình luận dưới dạng văn bản, ẩn chứa trong đó là cảm xúc của khách hàng liên quan tới chất lượng của các sản phẩm mà họ quan tâm. Nghiên cứu này đề xuất một mô hình kiến trúc hệ thống hỗ trợ khách hàng ra quyết định mua hàng trực tuyến dựa trên phương pháp khai thác dữ liệu phi cấu trúc. Dữ liệu nghiên cứu được thu thập trên các trang thương mại điện tử lớn của Việt Nam, sau đó được phân loại thành tích cực hoặc tiêu cực bởi các mô hình của phương pháp học máy có giám sát. Sau khi thử nghiệm và đánh giá, chúng tôi đã lựa chọn mô hình Support Vector Machine (SVM) có độ chính xác cao nhất để làm mô hình thực nghiệm. Nghiên cứu có giá trị tham khảo cho các nhà nghiên cứu trong lĩnh vực thương mại điện tử và các lĩnh vực khác của kinh doanh và quản lý.

Từ khóa - Khai thác dữ liệu phi cấu trúc; học máy có giám sát; hệ hỗ trợ ra quyết định mua hàng; mô hình phân loại cảm xúc

Abstract - The types of unstructured data left behind by customers in cyberspace are becoming more important for online businesses. That type of unstructured data is textual comments, containing feelings of customers related to the quality of the items which they are interested in. This study aims to propose a system architecture model to support customers in making online purchasing decisions based on the unstructured data mining. Research data are customers' comments collected on major Vietnamese e-commerce websites, and then classified into positive or negative by models of Supervised Machine Learning methods. After testing and evaluated, we selected the Support Vector Machine (SVM) model with the highest accuracy to make the experimental model. The study is of reference value for researchers in the field of e-commerce and other fields of business and management.

Key words - Unstructured data mining; supervised machine learning; purchase decision support system; sentiment classification model

1. Giới thiệu

Mua sắm thông qua nền tảng thương mại điện tử đã trở thành xu hướng tất yếu trong thời đại hiện nay. Đặc biệt trong bối cảnh bị ảnh hưởng bởi dịch bệnh Covid-19 thì số lượng người tham gia mua sắm trên các nền tảng thương mại điện tử đã ra tăng một cách nhanh chóng. Khi một khách hàng sau khi trải nhiệm dịch vụ mua sắm trên một trang thương mại điện tử, hoặc đã từng sử dụng sản phẩm được bán trên trang đó thì thường sẽ để lại những đánh giá, bình luận thông qua chức năng tự động của hệ thống [1]. Những bình luận dạng văn bản như vậy còn gọi là dữ liệu phi cấu trúc. Ở khía cạnh người mua hàng tiếp theo, sau khi họ quan tâm tới một mặt hàng nào đó, thường có xu hướng truy cập vào các trang thương mại điện tử để xem và trải nghiệm trước mặt hàng, hoặc tham khảo các bình luận của những khách hàng trước, sau đó mới đưa ra quyết định có mua hay không [2]. Dữ liệu bình luận được tạo ra bởi khách hàng đang gia tăng không ngừng trên các hệ thống website theo thời gian thực. Đây là một nguồn tài nguyên dữ liệu rất quan trọng cho các doanh nghiệp để có thể nhận biết tâm lý, xu hướng của khách hàng, từ đó cải thiện chất lượng dịch vụ khách hàng, hỗ trợ mua hàng để tăng doanh thu. Tuy nhiên, làm thế nào để khai thác được dữ liệu này một cách hiệu quả mà không mất nhiều thời gian, chi phí nguồn lực? Và áp dụng như thế nào để hỗ trợ khách hàng lựa chọn sản phẩm, giúp nâng cao hiệu quả bán hàng? Xuất phát từ những vấn đề trên, nhóm tác giả hình thành ý tưởng khai thác những bình luận này bởi chương trình máy tính tự động và thực hiện phân loại bởi phương pháp học máy nhằm hỗ trợ khách hàng ra quyết định lựa chọn sản phẩm trong mua sắm trực tuyến.

2. Các nghiên cứu liên quan

Nghiên cứu về hỗ trợ khách hàng mua hàng trực tuyến đã được nhiều tác giả quan tâm. Đặc biệt, trong vài năm trở lại đây, từ khi internet tăng tốc và thương mại điện tử phát triển mạnh mẽ, đã có nhiều các mô hình hỗ trợ khách hàng mua hàng trực tuyến được đề xuất như: Mô hình hệ thống hỗ trợ mua hàng dựa vào thông tin nhân khẩu học, hệ thống này thực hiện điều chuyển người dùng tới website bán hàng phù hợp dựa vào các thông tin được thu thập từ khách hàng, như thông tin địa lý hay độ tuổi [3, 4]; Mô hình hỗ trợ dựa trên lý thuyết giá trị đa thuộc tính (MAVT), hỗ trợ dựa trên thông tin mô tả mặt hàng cùng với sở thích của khách hàng [5]. Mô hình dựa vào sự tương tác của khách hàng với sản phẩm trong quá khứ để hỗ trợ lựa chọn mặt hàng tương tự [6]. Hoặc mô hình dựa vào cùng sở thích với khách hàng khác để hỗ trợ lựa chọn mặt hàng [7]; Mô hình dựa vào độ tương đồng giữa các mặt hàng trong cùng hệ thống [8-9].

Nhìn chung, các mô hình thu thập được chỉ dựa vào thông tin nhân khẩu học của khách hàng, dựa vào mối quan hệ cơ học giữa khách hàng với mặt hàng, và sự liên quan của các sản phẩm trong cùng hệ thống để hỗ trợ khách hàng

[1] Thai Nguyen University - University of Information and Communication Technology (Le Trieu Tuan)
[2] National Economics University (Pham Minh Hoan)

lựa chọn sản phẩm. Việc phân tích dữ liệu bị giới hạn trong một miền nhất định, phụ thuộc vào các mối quan hệ của các đối tượng khách hàng, sản phẩm trong quá khứ và không đưa ra được cái nhìn sâu sắc về xu hướng và sự vận động của sự hài lòng đến từ khách hàng. Điều này có thể gây ra sự lưỡng lự trong việc đưa ra quyết định lựa chọn sản phẩm của khách hàng. Bên cạnh đó, những phương pháp này không thể giám sát sự hài lòng của khách hàng một cách liên tục, và không có khả năng theo dõi xu hướng hài lòng của khách hàng trong dài hạn [10].

Trong nước, cũng bắt đầu có những nghiên cứu sử dụng phương pháp liên quan tới phân tích dữ liệu phi cấu trúc để hỗ trợ khách hàng trực tuyến. Điển hình là nghiên cứu [11] đã tiến hành thực nghiệm việc phân loại các bình luận trên bộ dữ liệu trong lĩnh vực thực phẩm bởi các mô hình thuật toán của phương pháp học máy như: Decision Tree, Naïve Bayes, hồi quy Logistic. Ngoài ra, còn có các nghiên cứu trong lĩnh vực du lịch [12]; nghiên cứu [13] sử dụng các mô hình Naive Bayes, Support Vector Machines và Maximum Entropy để phân loại các bình luận về khách sạn tại Việt Nam; Nghiên cứu so sánh các phương pháp phân loại bình luận bằng Tiếng Việt [14].

Hiện nay, với sự bùng nổ của dữ liệu lớn (Big Data), cách thức tương tác của khách hàng với các nền tảng bán hàng cũng đã dần thay đổi. Kéo theo đó là sự cần thiết phải thay đổi cách thức tiếp cận trong việc hỗ trợ khách hàng mua hàng của doanh nghiệp hay các nhà quan tâm. Và các mô hình hệ thống cũng cần thay đổi theo hướng sử dụng dữ liệu lớn [15]. Nghiên cứu này khác so với những nghiên cứu trên ở chỗ, nhóm tác giả khai thác dữ liệu phi cấu trúc; Cụ thể là những bình luận dạng văn bản của khách hàng để nhận biết những cảm nhận tích cực hay tiêu cực trên từng sản phẩm, qua đó cung cấp thông tin hỗ trợ khách hàng ra quyết định lựa chọn sản phẩm.

3. Cơ sở lý thuyết

3.1. Ra quyết định và hỗ trợ ra quyết định trong mua hàng trực tuyến

Quyết định mua hàng là mô hình hành vi của người tiêu dùng tuân theo một quy trình ra quyết định bao gồm các giai đoạn khác nhau để đạt được sự lựa chọn [16]. Mỗi người có những cách mua khác nhau đối với bất kỳ một sản phẩm nhất định nào, nghiên cứu [17] cho rằng, khách hàng đã quen với việc thay đổi cách tiếp cận ra quyết định theo các môi trường và tính huống khác nhau, và luôn cố gắng giảm thiểu nỗ lực liên quan tới nhận thức. Và trong trường hợp này, họ thường tìm kiếm sự hỗ trợ khi họ gặp phải quá nhiều thông tin để ít tốn công sức và thời gian hơn trong việc đưa ra quyết định tốt hơn [18]. Ngày nay, do sự phổ biến của thương mại điện tử, khi tìm hiểu thông tin mặt hàng khách hàng thường tìm đọc những nhận xét, đánh giá của những khách hàng trước về sản phẩm đó [19]. Số lượng mặt hàng trên các website thường là rất lớn và đa dạng, người tiêu dùng thường không thể đánh giá sâu được hết các sản phẩn lựa chọn có sẵn trên đó [20] và ở giai đoạn đầu tiên họ thường lọc ra một tập hợp các sản phẩm, sau đó xác định các sản phẩm hứa hẹn nhất [21]. Những sản phẩm được lựa chọn có xu hướng ảnh hưởng bởi các đánh giá tích cực hay tiêu cực của những người dùng trước [22]. Khai thác lượng dữ liệu phi cấu trúc khổng lồ được tạo ra trong quá trình giao dịch để hiểu sâu sắc hơn về hành vi khách hàng là rất cần thiết để hỗ trợ người mua hàng [23].

Hệ hỗ trợ ra quyết định (Decision Support System – DSS) là hệ thống thông tin dựa trên máy tính có thể hỗ trợ việc ra quyết định bằng cách phân tích dữ liệu và cung cấp thông tin cho người dùng [23]. Các DSS áp dụng các công cụ giúp người tiêu dùng lựa chọn sản phẩm có thể ảnh hưởng phần lớn đến việc ra quyết định của họ [24] và có tác động lớn tới tất cả các loại quyết định trong kinh doanh [25]. Có hai cách tiếp cận để phát triển DSS hỗ trợ người tiêu dùng trực tuyến đó là tiếp cận theo hướng dữ liệu [26] và tiếp cận theo hướng tri thức [27].

3.2. Khai thác dữ liệu phi cấu trúc

3.2.1. Khai thác văn bản

Dữ liệu phi cấu trúc thường đề cập đến những thông tin không được định nghĩa trước về mô hình dữ liệu quan hệ [28]. Hiện nay, trên các hệ thống kinh doanh trực tuyến, hơn 80% dữ liệu tồn tại ở dạng này [29], trong đó phổ biến và hữu ích nhất là dạng văn bản [30] được tạo ra từ những đánh giá sản phẩm của khách hàng. Những dòng văn bản đánh giá có thể được đọc hiểu, phân tích để thu được những thông tin kinh doanh một cách thủ công. Tuy nhiên với một lượng lớn dữ liệu thì cách xử lý này sẽ không hiệu quả. Công nghệ Big Data và kỹ thuật xử lý ngôn ngữ tự nhiên phát triển cho phép khai thác những dạng dữ liệu này theo những quy trình tự động.

Khai thác văn bản là quá trình trích xuất thông tin hữu ích và ý nghĩa từ văn bản [31]. Các phương pháp, công cụ khai thác dữ liệu có thể giúp khám phá kiến thức ẩn trong các nội dung văn bản của khách hàng và giúp doanh nghiệp hiểu khách hàng theo cách tốt hơn [32]. Học máy kết hợp với xử lý ngôn ngữ tự nhiên là kỹ thuật khai thác phổ biến và khả thi nhất hiện nay. Nó có thể giúp phân loại dữ liệu văn bản thành các danh mục khác nhau, để hiểu xu hướng hoặc chuyển động của dữ liệu, phát hiện sự giống nhau trong các tập dữ liệu và dự đoán tương lai dựa trên quá khứ [33].

Thông tin có sẵn ở dạng văn bản được chia thành hai phần, khách quan (objective) và chủ quan (subjective). Các sự kiện có thể được thể hiện bằng các nội dung khách quan, trong khi nhận thức, quan điểm tình cảm được thể hiện ở các khía cạnh chủ quan. Trong xử lý ngôn ngữ tự nhiên, trọng tâm là khai thác thông tin thực tế từ văn bản, tức thông tin dưới dạng khách quan. Tuy nhiên, với sự phát triển của công nghệ web, công nghệ khai thác Big Data giúp khai thác kiến thức nội dung do người dùng tạo ra, đây được gọi là phân tích chủ quan, hay phân tích tình cảm [34].

3.2.2. Phân tích tình cảm

Bình luận của khách hàng chứa những tình cảm và trải nghiệm của họ liên quan tới sản phẩm, dịch vụ [35-37]. Dữ liệu đánh giá, bình luận sản phẩm là một giải pháp để thu thập dữ liệu, nó cung cấp thông tin hữu ích cho nhà quản lý, ảnh hưởng đến hành vi mua hàng của khách hàng [38, 39] và cả hoạt động của công ty [40]. Vì vậy, các nhà quản lý có thể trích xuất những thông tin chi tiết có giá trị như vậy từ dữ liệu đánh giá, bình luận và hành động theo đó. Nội dung đánh giá, bình luận trực tuyến của các khách hàng về các mặt hàng là một nguồn thông tin phong phú, được coi là một gợi ý thân thiện giữa các khách hàng [41].

Tình cảm của khách hàng trong các bình luận gồm có

trạng thái tích cực và tiêu cực [42], phân tích tình cảm tức là phân loại văn bản theo hướng tích cực hoặc tiêu cực [43, 44]. Theo các nghiên cứu [45, 46] phân tích tình cảm cực kỳ hữu ích trong việc hỗ trợ khách hàng ra quyết định, giúp các nhà quản lý hiểu được sở thích của khách hàng, theo dõi và giám sát sự vận động xu hướng mong muốn về sản phẩm hoặc dịch vụ của họ.

3.2.3. Kỹ thuật xác định độ quan trọng của từ

Trong nghiên cứu này, độ quan trọng của từ được xác định bởi phương pháp TF-IDF (Term Frequency – Inverse Document Frequency) [47]. Là một kỹ thuật được sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản. Giá trị TF-IDF của từ khóa w_i trong bình luận d được tính bằng công thức sau:

$$\text{Tf\_idf} = \text{tf(wi, dj)} \times \log\frac{N}{n_i} \quad (1)$$

Trong đó:

$\text{tf}(w_i, d_j)$: Tần suất xuất hiện của từ khóa w_i trong văn bản dj.

$$F_{id} = \frac{\text{số lần } w_i \text{ xuất hiện trong văn bản } d_j}{\text{tổng số từ trong văn bản } d_j} \quad (2)$$

N: Tổng số văn bản trong tập mẫu;

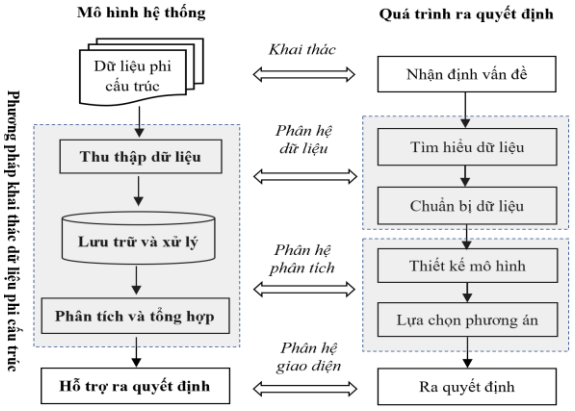
ni: Số văn bản có từ khóa wi.

4. Phương pháp nghiên cứu

Nghiên cứu này, nhóm tác giả sử dụng phương pháp nghiên cứu định lượng, các dữ liệu được thu thập trực tiếp từ trang thương mại điện tử. Sau đó, phương pháp học máy có giám sát (Suppervised Machine Leanring) được sử dụng để phân tích và tổng hợp dữ liệu. Quá trình ra quyết định thực hiện theo chuẩn công nghiệp CRIP-DM (Cross Industry Standard Process for Data Mining) bao gồm các bước [48]: Nhận định vấn đề; Tìm hiểu dữ liệu; Chuẩn bị dữ liệu; Thiết kế mô hình; Lựa chọn phương án; Ra quyết định. Môi trường thực nghiệm nghiên cứu được cài đặt bằng ngôn ngữ lập trình Python với sự hỗ trợ của công cụ tách từ Underthesea dành cho ngôn ngữ Tiếng Việt và các thư viện có sẵn.

5. Mô hình nghiên cứu đề xuất

Xuất phát từ cơ sở lý thuyết và các công trình nghiên cứu liên quan, mô hình nghiên cứu tổng quát được đề xuất như Hình 1.



Hình 1. *Mô hình nghiên cứu tổng quát*

5.1. Khai thác dữ liệu phi cấu trúc

5.1.1. Thu thập dữ liệu

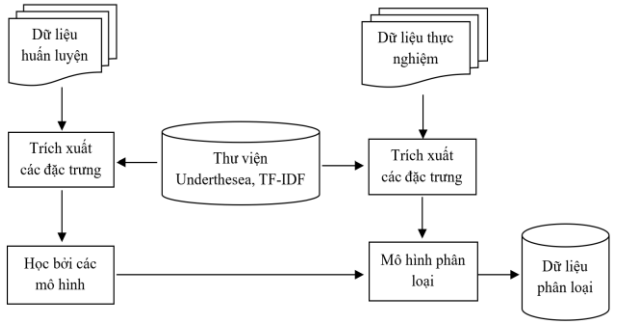
Dữ liệu bình luận bằng Tiếng Việt được thu thập từ một số trang thương mại điện tử hàng đầu tại Việt Nam bởi chương trình máy tính tự động Selenium Python. Đây là phương pháp thu thập nội dung dựa vào cấu trúc Hypertext Markup Language (HTML) của các trang web [49].

5.1.2. Lưu trữ và xử lý

Dữ liệu thu thập được lưu trữ ở định dạng CSV. Tiếp đến, nghiên cứu đã tiến hành tiền xử lý dữ liệu bằng cách loại bỏ những bình luận bị khuyết, những câu không ý nghĩa, câu không phải Tiếng Việt, dấu chấm, dấu phẩy dư thừa, những phản hồi không chứa đựng thông tin cần thiết... Tách câu thành các từ hoặc từ ghép có nghĩa bằng thư viện Underthesea [50] và chuyển đổi dữ liệu văn bản thành vector bằng phương pháp TF-IDF. Bộ dữ liệu dùng để thử nghiệm sẽ được chia theo tỷ lệ 80% dành cho huấn luyện (training) và 20% dành cho thử nghiệm (testing). Thực hiện gán nhãn (phân loại) dữ liệu theo phương pháp của [51] dựa vào điểm số đánh giá (rating) của khách hàng. Sau khi xem xét ngẫu nhiên tập dữ liệu thu thập, chúng tôi nhận thấy những bình luận có điểm số *rating >= 3* là tích cực (positive) và ngược lại *rating < 3* là tiêu cực (negative). Chúng tôi không xét những bình luận trung tính (neutral) do chúng không có ý nghĩa để khuyến nghị. Những dòng bình luận không được đánh giá điểm số, chúng tôi sẽ thực hiện gán nhãn thủ công.

5.1.3. Phân tích và tổng hợp

Quá trình phân loại và tổng hợp kết quả phân loại dữ liệu được mô tả như Hình 2.



Hình 2. *Mô hình hệ thống phân loại dữ liệu*

Giai đoạn này nhằm, các mô hình của hoc máy có giám sát sẽ được huấn luyện, bao gồm: mô hình Support Vector Machine (SVM), Naive Bayes (NB), Random Forrest (RF), Neural Network (NN) và Decision Tree (DT). Sau đó thử nghiệm, đánh giá và lựa chọn ra mô hình có độ chính xác cao nhất để thực nghiệm.

Bảng 1. *Ma trận nhầm lẫn*

	Thực tế: positive	**Thực tế: negative**
Dự đoán: positive	True Positive (TP)	False Negative (FN)
Dự đoán: negative	False Positive (FP)	True Negative (TN)

Nguồn: [52]

Nghiên cứu dùng phương pháp đánh giá mô hình phổ biến là dựa trên các chỉ số tính toán trong ma trận nhầm lẫn (Confusion Matrix). Hiệu quả của mô hình được đánh giá dựa trên 4 chỉ số: Độ chính xác (Accuracy); Độ hội tụ

(Precision); Độ bao phủ (Recall) và Giá trị trung bình điều hòa (F1-score) cho biết hiệu quả tổng thể, F1-score có giá trị càng cao thì mô hình phân loại càng chính xác.

Trong đó:

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN} \quad (3)$$

$$\text{Precesion} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

True Positive (TP): Tổng số lượng bình luận tích cực dự đoán Đúng so với thực tế.

False Positive (FP): Tổng số lượng bình luận tích cực dự đoán Sai so với thực tế.

True Negative (TN): Tổng số lượng bình luận tiêu cực dự đoán Đúng so với thực tế.

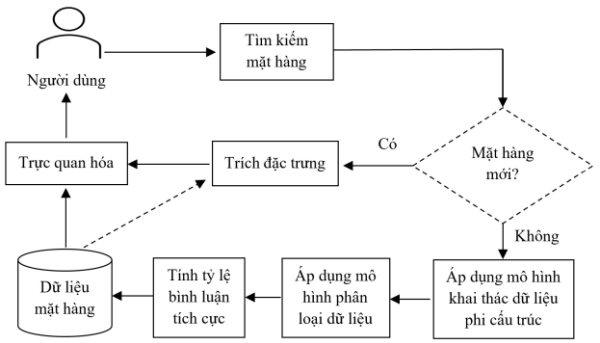
False Negative (FN): Tổng số lượng bình luận tiêu cực dự đoán Sai so với thực tế.

5.2. Hỗ trợ ra quyết định

Để hỗ trợ cho khách hàng ra quyết định khi mua hàng, dữ liệu bình luận về mặt hàng R mà khách hàng quang tâm được đưa vào mô hình để phân loại. Kết quả tỷ lệ bình luận tích cực (R_{pos}) được tính và hiện thị cung cấp thông tin cho khách hàng ra quyết định lựa chọn.

$$R_{pos} = \frac{Pos}{\sum N_i} \quad (7)$$

Trong đó: *Pos* là số lượng bình luận tích cực, N_i là bình luận thứ *i* trên mặt hàng R.



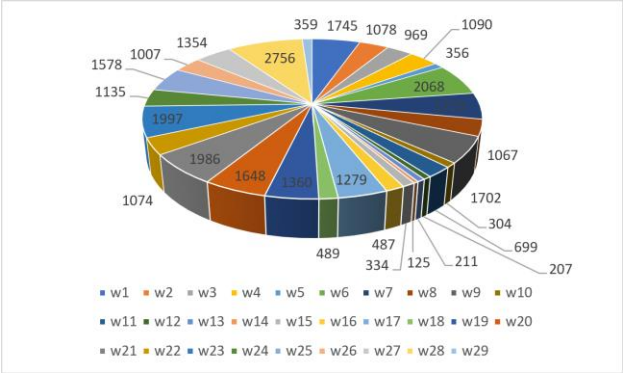
Hình 3. *Mô hình hệ hỗ trợ khách hàng ra quyết định lựa chọn mặt hàng*

Hình 3 là mô hình hỗ trợ ra quyết định cho khách hàng lựa chọn mặt hàng. Đầu tiên, khách hàng tìm kiếm mặt hàng cần mua, nếu mặt hàng đó đã được những khách hàng khác đánh giá (mặt hàng cũ) thì thực hiện hiện áp dụng mô hình khai thác dữ liệu phi cấu trúc để thu thập và phân loại các bình luận, sau đó tính tỷ lệ bình luận tích cực, lưu vào cơ sở dữ liệu mặt hàng và tổ chức hiển thị kết quả tới người dùng. Trong trường hợp mặt hàng đó chưa có người dùng nào đánh giá (mặt hàng mới) thì trích những đặc trưng liên quan tới mặt hàng đó từ cơ sở dữ liệu và tổ chức hiển thị tới người dùng.

6. Kết quả

6.1. Kết quả thu thập và tiền xử lý dữ liệu

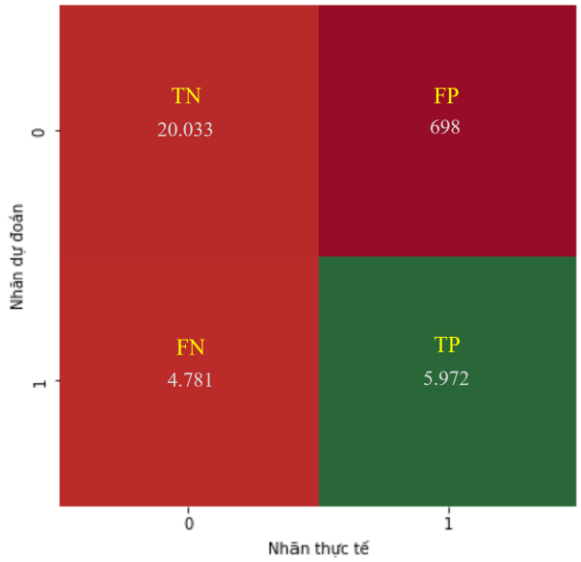
Nghiên cứu đã tiến hành thu thập tự động được 33.417 bình luận từ năm 2017 đến 2022 trên 29 website thương mại điện tử hàng đầu tại Việt Nam. Sau khi xử lý, loại bỏ những bình luận không liên quan, bị lỗi phông chữ, những câu không ý nghĩa, dữ liệu còn lại để thực nghiệm là 32.187 bình luận được phân bố như trong Hình 4. Tập dữ liệu này được chia thành tập dữ liệu dùng cho huấn luyện, thực hiện gán nhãn và tập dữ liệu dành cho thử nghiệm.



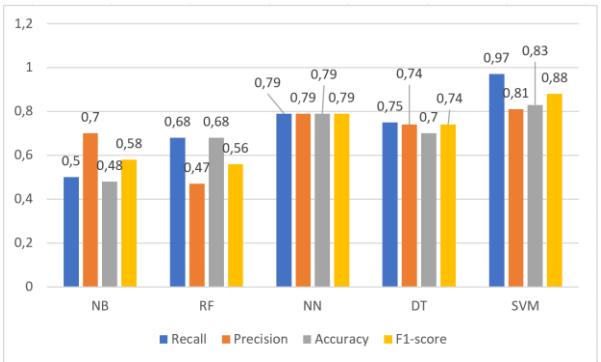
Hình 4. *Phân bố số lượng các bình luận trên các website (w)*

6.2. Kết quả huấn luyện mô hình

Kết quả huấn luyện các mô hình được thể hiện trong Hình 6.



Hình 5. *Ma trận nhầm lẫn của mô hình SVM*



Hình 6. *Kết quả huấn luyện các mô hình*

Kết quả huấn luyện cho thấy mô hình SVM có độ chính xác cao nhất (88%), do đó mô hình này sẽ được lựa chọn để áp dụng cho dữ liệu thực nghiệm.

6.3. Trực quan hóa hỗ trợ ra quyết định mua hàng

Việc áp dụng mô hình sẽ giúp khách hàng biết được mặt hàng có ý định mua trên một website thương mại điện tử cụ thể có được nhiều khách hàng trước đó đánh giá tích cực hay không. Kết hợp với dữ liệu phân loại bình luận của các khách hàng trước sẽ giúp khách hàng dễ dàng đưa ra quyết định mua hàng.

Bảng 2. *Kết quả hỗ trợ ra quyết định mua của một số mặt hàng*

Mặt hàng quan tâm	Trên hệ thống website	Tỷ lệ bình luận tích cực			
		Tổng	Tích cực	Tiêu cực	Tỷ lệ
Tivi Samsung	W26	52	32	4	61%
Tủ lạnh Panasonic	W4	27	24	2	90%
Điều hòa Casper	W14	13	12	0	94%
Quần Jean	W1	45	45	0	100%
Áo thun	W10	156	137	10	88%
Quần bơi nam	W5	34	31	1	90%
Điện thoại Iphone 12	W3	29	25	2	86%
Điện thoại Iphone 11 pro max	W8	20	19	1	94%
Áo chống nắng	W6	75	75	0	100%
Gà quay chiên ròn	W21	61	34	22	56%
Sản phẩm giúp giảm cân	W11	33	27	4	81%
Sườn dim xì dầu	W21	18	7	10	38%

6.4. Thảo luận kết quả nghiên cứu

Từ kết quả nghiên cứu, bài báo đưa ra một số thảo luận dựa trên một số khía cạnh để có thể triển khai hệ thống vào thực tiễn lĩnh vực thương mại điện tử như sau:

Phạm vi triển khai hệ thống: Với đặc tính của hệ thống sử dụng nguồn dữ liệu thứ cấp sẵn có nên đề xuất cho doanh nghiệp có thể xây dựng và triển khai một hệ thống độc lập, thực hiện khai thác nguồn dữ liệu trên mạng để đánh giá chất lượng dịch vụ khách hàng của các hệ thống website thương mại điện tử phục vụ công tác quản lý và thực hiện hỗ trợ khách hàng mua hàng.

Công nghệ lưu trữ và xử lý dữ liệu: Hiệu xuất xử lý của hệ thống và khả năng hỗ trợ nhà quản lý, khách hàng ra quyết định phụ thuộc lớn vào độ lớn của tập dữ liệu và năng lực xử lý của hệ thống máy tính. Do đó, khi triển khai thực tế, doanh nghiệp cũng cần tính toán đến công nghệ lưu trữ dữ liệu lớn.

Hệ thống có dữ liệu đầu vào lớn và đòi hỏi xử lý phức tạp, mất nhiều thời gian. Do đó, chức năng thu thập, tiền xử lý dữ liệu, huấn luyện lại các mô hình nên được thực hiện theo định kỳ. Bên cạnh đó, cũng tùy thuộc vào tốc độ tăng trưởng và biến động của nguồn dữ liệu bình luận của khách hàng trên các website thương mại điện tử.

Bên cạnh đối tượng sử dụng chính của hệ thống là nhà quản lý, quản trị doanh nghiệp và khách hàng thì các chức năng thu thập, tiền xử lý dữ liệu, huấn luyện, đánh giá và lựa chọn các mô hình nên được thực hiện bởi các chuyên gia tri thức, đặc biệt là các chuyên gia về khoa học dữ liệu.

7. Kết luận

Nghiên cứu đã đề xuất một mô hình hỗ trợ người mua hàng ra quyết định mua dựa trên phân tích dữ liệu phi cấu trúc là các bình luận của khách hàng trên các website thương mại điện tử. Các mô hình phân loại của phương pháp học máy được huấn luyện, thử nghiệm, đánh giá và đã lựa chọn ra mô hình SVM có độ chính xác cao nhất làm mô hình thực nghiệm. Khách hàng quan tâm tới bất kỳ sản phẩm nào trên website được triển khai hệ thống sẽ không phải đọc hiểu các bình luận thủ công, hệ thống sẽ phân loại các bình luận một cách nhanh chóng và hiển thị cho khách hàng. Tuy nhiên, nghiên cứu vẫn còn một số hạn chế có thể cải thiện tốt hơn ở các nghiên cứu tiếp theo. Hạn chế về đối tượng và phạm vi nghiên cứu: Nghiên cứu chỉ thực hiện thu thập dữ liệu ở dạng tĩnh, mà thực tế quyết định mua hàng của khách hàng còn phụ thuộc vào những yếu tố khách quan khác, như vị trí địa lý của công ty, sở thích, đặc trưng văn hóa vùng miền. Bên cạnh đó, hệ thống chưa thực hiện thu thập dữ liệu trên toàn bộ hệ thống website thương mại điện tử tại Việt Nam, đồng thời chỉ thực hiện xử lý trên ngôn ngữ Tiếng Việt, hệ thống có thể mở rộng sang các dạng ngôn ngữ khác; Hạn chế về phương pháp nghiên cứu: Nghiên cứu chỉ phân loại nội dung bình luận theo thang đo hai mức tích cực và tiêu cực. Hướng nghiên cứu tiếp theo có thể sử dụng thang đo nhiều mức hơn (ví dụ thang đo Likert 5 mức). Bên cạnh đó, nghiên cứu chỉ sử dụng phương pháp phân loại học máy có giá sát, nếu kết hợp thêm phương pháp lọc nội dung và phương pháp từ vựng dựa trên ngữ nghĩa thì có thể sẽ cho kết quả tốt hơn.

TÀI LIỆU THAM KHẢO

[1] Mudambi, S. and D. Schuff, "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com", *MIS Quarterly*, 34, 2010, 185-200.

[2] Sharma, D.K., et al., "E-Commerce product comparison portal for classification of customer data based on data mining", *Materials Today: Proceedings*, 51, 2022, 166-171.

[3] Al-Shamri, M.Y.H., "User profiling approaches for demographic recommender systems", *Knowledge-Based Systems*, 100, 2016, 175-187.

[4] Xu, J., Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view", *Information Sciences*, 505, 2020, 772-794.

[5] Pazzani, M.J. and D. Billsus, *Content-based recommendation systems*, in *The adaptive web*, Springer, 2007, 325-341.

[6] Patra, B.G., et al., "A content-based literature recommendation system for datasets to improve data reusability – A case study on Gene Expression Omnibus (GEO) datasets", *Journal of Biomedical Informatics*, 104, 2020, 1-14.

[7] Afoudi, Y., M. Lazaar, and M. Al Achhab, "Impact of Feature selection on content-based recommendation system", *International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 2019, 1-6.

[8] Aljunid, M.F. and M. Dh, "An Efficient Deep Learning Approach for Collaborative Filtering Recommender System", *Procedia Computer Science*, 171, 2020, 829-836.

[9] Ghasemi, N. and S. Momtazi, "Neural text similarity of user reviews for improving collaborative filtering recommender systems", *Electronic Commerce Research and Applications*, 45, 2021, 101019.

[10] Zhang, F., et al., "Graph embedding-based approach for detecting group shilling attacks in collaborative recommender systems", *Knowledge-Based Systems*, 199(7), 2020, 105984.

[11] Yussupova, N., et al., "Models and Methods for Quality Management Based on Artificial Intelligence Applications", *Acta Polytechnica Hungarica*, 13(3), 2016, 45-60.

[12] Nguyễn Đặng Lập Bằng, Nguyễn Văn Hồ, & Hồ Trung Thành, "Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm", *Tạp chí Khoa học Đại học Mở Thành phố Hồ Chí Minh*, 16(1), 2020, 64-78.

[13] Duyen, N.T., N.X. Bach, and T.M. Phuong, "An empirical study on sentiment analysis for Vietnamese", in *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, 2014, IEEE.

[14] Thái Kim Phụng, Nguyễn An Tế, & Trần Thị Thu Hà, "Tiếp cận phương pháp học máy trong khai thác ý kiến khách hàng trực tuyến", *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á*, 30(10), 2019, 27-41.

[15] Bang, T.S., C. Haruechaiyasak, and V. Sornlertlamvanich, "Vietnamese sentiment analysis based on term feature selection approach", in *Proc. 10th International Conference on Knowledge Information and Creativity Support Systems (KICSS 2015)*, 2015.

[16] Darley, W., Blankson, C., & Luethge, D., "Toward an Integrated Framework for Online Consumer Behavior and Decision Making Process: A Review", *Psychology and Marketing*, 27(2), 2010, 94-116.

[17] Shugan, S.M., "The Cost Of Thinking", *Journal of Consumer Research*, 7(2), 1980, 99-111.

[18] Payne, J.W.J.P.b., "Contingent decision behavior", *Psychological Bulletin*, 92(2), 1982, 382-402.

[19] Häubl, G. and V.J.M.s. Trifts, "Consumer decision making in online shopping environments: The effects of interactive decision aids", *Marketing Science*, 19(1), 2000, 4-21.

[20] Bhargave, R., A. Chakravarti, and A. Guha, "Two-Stage Decisions Increase Preference for Hedonic Options", *Organizational Behavior and Human Decision Processes*, 130, 2015, 123-135.

[21] Yang, L., M. Xu, and L. Xing, "Exploring the core factors of online purchase decisions by building an E-Commerce network evolution model", *Journal of Retailing and Consumer Services*, 64, 2022, 102784.

[22] Kart, Ö., A. Kut, and V. Radevski, "Decision Support System For A Customer Relationship Management Case Study", *International Journal of Informatics and Communication Technology (IJ-ICT)*, 3, 2014, 88-96.

[23] Bharati, P. and A.J.D.s.s. Chaudhury, "An empirical investigation of decision-making satisfaction in web-based decision support systems", *Decision Support System*, 37(2), 2004, 187-197.

[24] Manivannan, S., "Application of Decision Support System in E-commerce", *Communications of the IBIMA*, 15, 2008, 156-169.

[25] Kasper, G.M., "A Theory of Decision Support System Design for User Calibration", *Information Systems Research*, 7(2), 1996, 215-232.

[26] Chandra, Y., S. Karya, and M. Hendrawaty, "Decision Support Systems for Customer to Buy Products with an Integration of Reviews and Comments from Marketplace E-Commerce Sites in Indonesia: A Proposed Model", *International Journal on Advanced Science, Engineering and Information Technology*, 9(4), 2019, 1171-1176.

[27] Jain, S., A. de Buitléir, and E. Fallon, "A Review of Unstructured Data Analysis and Parsing Methods", *IEEE International Conference on Emerging Smart Computing and Informatics (IEEE – ESCI 2020)*, Web of Science Journal Publication, 2020.

[28] He, P., et al., "An Evaluation Study on Log Parsing and Its Use in Log Mining", in *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2016.

[29] Inmon, W.H. and D. Linstedt, *2.4 - Unstructured Data*, in *Data Architecture: a Primer for the Data Scientist*, W.H. Inmon and D. Linstedt, Editors, Morgan Kaufmann: Boston, 2015, 63-70.

[30] Alzate, M., M. Arce-Urriza, and J., "Cebollada, Mining the text of online consumer reviews to analyze brand image and brand positioning", *Journal of Retailing and Consumer Services*, 67(1), 2022, 102989.

[31] Dahiya, A., N. Gautam, and P. Gautam, "Data Mining Methods and Techniques for Online Customer Review Analysis: A Literature Review", *Journal of System and Management Sciences*, 11(3), 2021, 1-26.

[32] Chen, J., et al., "Big data challenge: A data management perspective", *Frontiers of Computer Science*, 7, 2013, 157-164.

[33] Liu, B., *Web data mining: exploring hyperlinks, contents, and usage data*, Springer, 1, 2011.

[34] Archak, N., A. Ghose, and P. Ipeirotis, *Deriving the Pricing Power of Product Features by Mining Consumer Reviews*, NET Institute, Working Papers, 57, 2007.

[35] Decker, R. and M.J.I.J.o.R.i.M. Trusov, "Estimating aggregate consumer preferences from online product reviews", *International Journal of Research in Marketing*, 27(4), 2010, 293-307.

[36] Cai, Y., et al., "A deep recommendation model of cross-grained sentiments of user reviews and ratings", *Information Processing & Management*, 59(2), 2022, 102842.

[37] Li, M., et al., "Helpfulness of Online Product Reviews as Seen by Consumers: Source and Content Features", *International Journal of Electronic Commerce*, 17, 2013, 101-136.

[38] Tirunillai, S. and G. Tellis, "Does Online Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance", *Marketing Science*, 31(2), 2011, 198-215.

[39] Floyd, K., et al., "How Online Product Reviews Affect Retail Sales: A Meta-analysis", *Journal of Retailing*, 90(2), 2014, 217-232.

[40] East, R., K. Hammond, and W. Lomax, "Measuring the impact of positive and negative word of mouth on brand purchase probability", *International Journal of Research in Marketing*, 25(3), 2008, 215-224.

[41] Lutfullaeva, M., et al., "Optimization of Sentiment Analysis Methods for classifying text comments of bank customers", *IFAC-PapersOnLine*, 51(32), 2018, 55-60.

[42] Morinaga, S., et al., "Mining product reputations on the Web", *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, 341-349.

[43] Cruz, F.L., et al., "Building layered, multilingual sentiment lexicons at synset and lemma levels", *Expert Systems with Applications*, 41(13), 2014, 5984-5994.

[44] Bakshi, R.K., et al., "Opinion mining and sentiment analysis", *2016 3rd international conference on computing for sustainable global development (INDIACom)*, IEEE, 2016.

[45] Gensler, S., et al., "Listen to Your Customers: Insights into Brand Image Using Online Consumer-Generated Product Reviews", *International Journal of Electronic Commerce*, 20, 2016, 112-141.

[46] Heilig, L., R. Stahlbock, and S. Voss, *From Digitalization to Data-Driven Decision Making in Container Terminals*, Handbook of Terminal Planning, Springer, 2019, 125-154.

[47] Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., & Sidorov, G., "Unsupervised sentence representations as word information series: Revisiting TF–IDF", *Computer Speech & Language*, 56, 2019, 107-129.

[48] Lê Triệu Tuấn & Đàm Thị Phương Thảo, "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), 2022, 49-52.

[49] Anh, V., "Underthesea document", *Under the sea*, 2018, [Online] Available: https://underthesea.readthedocs.io, 02/10/2022.

[50] Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., & Sidorov, G., "Unsupervised sentence representations as word information series: Revisiting TF–IDF", *Computer Speech & Language*, 56, 2019, 107-129.

[51] Kulkarni, A., D. Chong, and F.A. Batarseh, *5 - Foundations of data imbalance and solutions for a data democracy*, in *Data Democracy*, F.A. Batarseh and R. Yang, Editors, Academic Press, 2020, 83-106.

[52] Sharma, D. K., Lohana, S., Arora, S., Dixit, A., Tiwari, M., & Tiwari, T., "E-Commerce product comparison portal for classification of customer data based on data mining", *Materials Today: Proceedings*, 51, 2022, 166-171.