ỨNG DỤNG MÔ HÌNH HỌC MÁY SVM TRONG PHÂN LOẠI CẢM XÚC TỪ BÌNH LUẬN KHÁCH HÀNG TRÊN AMAZON

APPLICATION OF SVM MACHINE LEARNING MODEL IN SENTIMENT CLASSIFICATION FROM CUSTOMER REVIEWS ON AMAZON

Nguyễn Thị Hồng Duyên[1,*], Vũ Thị Hồng Ánh[2]

DOI: https://doi.org/10.57001/huih5804.2025.235

TÓM TẮT

Bài báo này đề xuất ứng dụng phương pháp học máy, cụ thể là mô hình SVM (Support Vector Machine), để phân loại cảm xúc trong bình luận của khách hàng trên sàn thương mại điện tử Amazon trong giai đoạn từ quý I năm 1996 đến quý III năm 2023. Dữ liệu nghiên cứu được thu thập từ nguồn Amazon Review'23 và xử lý trên nền tảng Google Colaboratory. Quá trình phân tích bao gồm các bước tiền xử lý dữ liệu, trích xuất đặc trưng văn bản bằng phương pháp Bag-of-Words (BoW) và huấn luyện mô hình nhằm phân loại bình luận theo hai nhóm cảm xúc: tích cực và tiêu cực. Kết quả thực nghiệm cho thấy, mô hình SVM đạt hiệu suất cao trong nhiệm vụ phân loại cảm xúc. Những phát hiện từ nghiên cứu này không chỉ khẳng định tính hiệu quả của SVM trong xử lý ngôn ngữ tự nhiên (NLP) mà còn cung cấp cơ sở thực tiễn để các doanh nghiệp ứng dụng công nghệ học máy vào phân tích phản hồi khách hàng. Từ đó, hỗ trợ dự báo xu hướng, tối ưu chiến lược kinh doanh và nâng cao trải nghiệm khách hàng.

***Từ khoá**: Học máy, SVM, phân loại cảm xúc, Amazon, xử lý ngôn ngữ tự nhiên.*

ABSTRACT

This article proposes the application of machine learning, specifically the Support Vector Machine (SVM) model, for sentiment classification of customer reviews on the Amazon E-commerce platform during the period from the first quarter of 1996 to the third quarter 2023. The research dataset was collected from Amazon Review'23 and processed using the Google Colaboratory platform. The analytical procedure encompasses data preprocessing, text feature extraction using the Bag-of-Words (BoW) method, and model training to categorize reviews into two sentiment classes: positive and negative. Empirical results indicate that the SVM model achieves high performance in sentiment classification tasks. The findings not only confirm the efficacy of SVM in natural language processing (NLP) but also provide a practical foundation for businesses to integrate machine learning technologies into customer feedback analysis, thereby supporting trend forecasting, optimization of business strategies, and enhancement of customer experience.

***Keywords:** Machine Learning, SVM, sentiment classification, Amazon, textual data.*

[1]Trường kinh tế, Trường Đại học Công nghiệp Hà Nội

[2]Sinh viên chuyên ngành Phân tích dữ liệu kinh doanh, Trường Đại học Công nghiệp Hà Nội

[*]Email: nguyethihongduyen@haui.edu.vn

Ngày nhận bài: 01/3/2025

Ngày nhận bài sửa sau phản biện: 06/4/2025

Ngày chấp nhận đăng: 28/6/2025

1. GIỚI THIỆU

Trong những năm gần đây, thương mại điện tử đã phát triển mạnh mẽ và trở thành một phần không thể thiếu trong nền kinh tế toàn cầu. Đặc biệt, dưới tác động của đại dịch Covid-19, người tiêu dùng ngày càng ưa chuộng hình thức mua sắm trực tuyến do tính tiện lợi, tiết kiệm thời gian và đa dạng về sản phẩm. Nhận thấy tiềm năng, nhiều doanh nghiệp toàn cầu tập trung

chiến lược bán hàng, trong đó hệ thống đánh giá của khách hàng đóng vai trò cốt lõi trong phát triển kinh doanh. Việc thu thập và phân tích ý kiến của khách hàng đã mang lại nhiều lợi ích quan trọng trong chiến lược marketing đồng thời cũng hỗ trợ dự báo xu hướng thị trường tài chính [4]. Bình luận của khách hàng không chỉ giúp người tiêu dùng đưa ra quyết định mua sắm chính xác mà còn ảnh hưởng đến uy tín của sản phẩm cũng như danh tiếng của người bán. Tuy nhiên, lượng bình luận khổng lồ mỗi ngày đòi hỏi phương pháp phân tích dữ liệu tiên tiến để khai thác thông tin hữu ích. Trong khi đó, trí tuệ nhân tạo (AI) ngày càng chứng minh vai trò trong việc xử lý và phân tích dữ liệu lớn. Trong lĩnh vực phân tích ý kiến và đánh giá cảm xúc, việc sử dụng các phương pháp học máy giúp cải thiện đáng kể độ chính xác trong việc phân tích tình cảm của văn bản [16]. Các phương pháp phân tích cảm xúc mới dựa trên AI đã được giới thiệu với mục đích giúp cải thiện khả năng nhận diện ý kiến và cảm xúc tiềm ẩn trong các đánh giá của khách hàng [4]. Những nghiên cứu này đã mở ra con đường cho sự phát triển của mô hình học sâu và nâng cao khả năng xử lý ngôn ngữ tự nhiên.

Amazon là một trong những nền tảng thương mại điện tử lớn nhất và phát triển nhanh nhất trên thế giới, được thành lập vào năm 1994 bởi Jeff Bezos tại Seattle, Mỹ. Ban đầu, Amazon chỉ hoạt động như một cửa hàng trực tuyến bán sách, nhưng nhanh chóng mở rộng sang các sản phẩm và dịch vụ khác cũng như nhiều lĩnh vực khác. Sự phát triển của Amazon gắn liền với quá trình đầu tư mạnh mẽ vào công nghệ, đặc biệt là trong lĩnh vực trí tuệ nhân tạo (AI), dữ liệu lớn (big data) và điện toán đám mây [11]. Amazon sử dụng mô hình học sâu để dự đoán nhu cầu mua sắm của người dùng, giúp tăng tỷ lệ chuyển đổi và doanh thu [22]. Amazon Web Services (AWS), một nhánh của Amazon, đã nổi lên như một trong những dịch vụ điện toán đám mây hàng đầu thế giới, cung cấp hạ tầng cho nhiều doanh nghiệp lớn. Việc triển khai trí tuệ nhân tạo (AI) thông qua các dịch vụ như Amazon Lex, cho phép xây dựng và vận hành chatbot giọng nói, hỗ trợ xử lý ngôn ngữ tự nhiên và nhận dạng giọng nói tự động [19]. Nhìn chung Amazon không chỉ là một nền tảng thương mại điện tử, mà còn là một biểu tượng của sự đổi mới trong ngành công nghiệp công nghệ toàn cầu. Thành công của Amazon đến từ chính chiến lược đầu tư vào công nghệ và khả năng linh hoạt đáp ứng nhu cầu người tiêu dùng, tạo nên hệ sinh thái thương mại điện tử tích hợp.

Trong nghiên cứu này, nhóm tác giả đã ứng dụng mô hình máy vector hỗ trợ (Support Vector Machine - SVM), một thuật toán học máy có giám sát thuộc lĩnh vực trí tuệ nhân tạo, để phân loại mức độ hài lòng trong các đánh giá của khách hàng đối với sản phẩm trên sàn thương mại điện tử Amazon. Bằng cách khai thác dữ liệu phản hồi của người tiêu dùng và áp dụng phương pháp học máy, nghiên cứu góp phần mở rộng phạm vi ứng dụng của SVM trong thương mại điện tử. Kết quả nghiên cứu cung cấp cho doanh nghiệp và nhà bán lẻ những thông tin có giá trị nhằm nâng cao chất lượng sản phẩm, tối ưu hóa chiến lược kinh doanh và cải thiện trải nghiệm khách hàng thông qua các phương pháp phân tích dữ liệu tiên tiến.

2. TỔNG QUAN NGHIÊN CỨU

2.1. Dữ liệu lớn (Big Data)

Dữ liệu lớn (Big Data) được định nghĩa là tập hợp dữ liệu có kích thước rất lớn, tốc độ sinh dữ liệu nhanh và có sự đa dạng cao, đòi hỏi các công nghệ phân tích tiên tiến mới có thể xử lý một cách hiệu quả và tối ưu [9]. Ban đầu, khái niệm này được mô tả qua mô hình 3V gồm khối lượng (Volume), tốc độ (Velocity) và đa dạng (Variety) [12]. Tuy nhiên, mô hình 3V đã được mở rộng thành mô hình 5V bao gồm khối lượng (Volume) - lượng dữ liệu to lớn được tạo ra mỗi ngày từ nhiều nguồn khác nhau; tốc độ (Velocity) - Tốc độ sinh và truyền tải dữ liệu nhanh chóng; đa dạng (Variety) - Dữ liệu đến từ nhiều định dạng khác nhau như văn bản, hình ảnh, video, âm thanh; độ tin cậy (Veracity) - Chất lượng và tính chính xác của dữ liệu, đảm bảo dữ liệu có thể sử dụng được; giá trị (Value) - Giá trị mà dữ liệu mang lại khi được phân tích đúng cách [9]. Dữ liệu lớn có thể được phân loại thành: dữ liệu có cấu trúc (dữ liệu được lưu trữ trong các bảng cơ sở dữ liệu có tổ chức như SQL), dữ liệu bán cấu trúc (dữ liệu có một số yếu tố cấu trúc nhưng không tuân theo mô hình cố định), dữ liệu phi cấu trúc (dữ liệu không theo khuôn mẫu cụ thể, bao gồm văn bản, hình ảnh, video, âm thanh, email và bình luận khách hàng) [8]. Sự gia tăng mạnh mẽ về khối lượng dữ liệu trong ngành khoa học và công nghiệp đã làm nổi bật nhu cầu cấp thiết đối với các giải pháp lưu trữ, xử lý và khai thác dữ liệu một cách hiệu quả [6]. Cũng theo nghiên cứu, dữ liệu lớn cần được khai thác và phân tích với các công cụ tiên tiến và thường theo quy trình: thu thập dữ liệu, lưu trữ dữ liệu, xử lý dữ liệu, phân tích dữ liệu, hiển thị dữ liệu. Với sự phát triển nở rộ của thương mại điện tử, dữ liệu phản hồi của khách hàng phát sinh từ các nền tảng như Amazon có tốc độ tăng trưởng nhanh chóng trong

từng giây và chứa nhiều thông tin giá trị. Đây vừa là cơ hội vừa là thách thức để ứng dụng AI trong dự đoán xu hướng tiêu dùng, nâng cao trải nghiệm khách hàng và tối ưu chiến lược kinh doanh. Việc khai thác dữ liệu một cách hiệu quả giúp doanh nghiệp tạo ra mô hình kinh doanh tối ưu và bền vững hơn.

2.2. Xử lý ngôn ngữ tự nhiên NLP - Text Classification

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một lĩnh vực quan trọng của trí tuệ nhân tạo (AI), giúp máy tính hiểu và xử ngôn ngữ con người. Các phương pháp NLP hiện đại chủ yếu dựa vào học máy (Machine Learning) và học sâu (Deep Learning) để trích xuất thông tin hữu ích từ dữ liệu văn bản lớn. Dữ liệu lớn đóng vai trò quan trọng trong sự phát triển của NLP [8, 23]. Sau khi được thu thập, dữ liệu phi cấu trúc với ngôn ngữ tự nhiên cần đi qua các bước: tiền xử lý, trích xuất đặc trưng, xây dựng mô hình, đánh giá. Tiền xử lý dữ liệu nhằm loại bỏ nhiễu và chuyển đổi văn bản về dạng số hóa bao gồm các kỹ thuật như tách từ, loại bỏ từ dừng và rút gọn từ. Trích xuất đặc trưng là bước chuyển đổi văn bản thành dạng số hóa mà mô hình có thể hiểu và xử lý. Sau khi trích xuất đặc trưng, văn bản có thể được xử lý và phân tích bằng các mô hình khác nhau như học máy và học sâu. Một mô hình mạng nơ-ron tích chập (CNN) được đề xuất ở cấp ký tự để phân loại văn bản [25]. Mô hình này có thể học được các đặc trưng quan trọng của văn bản mà không cần dựa vào các bộ từ điển từ trước. Đây là một bước tiến lớn trong NLP, giúp cải thiện độ chính xác trong phân tích bình luận khách hàng. Bên cạnh đó, SVM - một mô hình học máy hiệu quả, đã khắc phục nhiều hạn chế của các phương pháp truyền thống, nâng cao độ chính xác trong phân loại và phân tích dữ liệu. Các chỉ số đánh giá cho thấy SVM có độ tin cậy cao và khả năng dự đoán chính xác.

Một trong những yếu tố then chốt trong phân tích bình luận của khách hàng là việc phát hiện và đánh giá cảm xúc. Một mô hình phân loại cảm xúc đã được phát triển vào năm 1980, bao gồm tám cảm xúc cơ bản như vui, buồn, tức giận, sợ hãi và các cảm xúc khác, tạo nền tảng lý thuyết quan trọng cho nghiên cứu này [17]. Các mô hình xử lý ngôn ngữ tự nhiên (NLP) hiện đại thường dựa vào lý thuyết này để xây dựng các hệ thống phân loại cảm xúc từ văn bản, giúp tự động nhận diện cảm xúc của người dùng trong các bình luận và đánh giá.

2.3. AI (Trí tuệ nhân tạo)

Trí tuệ nhân tạo (Artificial Intelligence - AI) là một lĩnh vực nghiên cứu liên ngành nhằm xây dựng các hệ thống có khả năng thực hiện các tác vụ yêu cầu trí tuệ của con người, bao gồm nhận diện ngôn ngữ, học tập và ra quyết định [18]. Trong bối cảnh thương mại điện tử, AI đóng vai trò quan trọng trong việc phân tích ý kiến khách hàng thông qua các phương pháp xử lý ngôn ngữ tự nhiên (NLP), học sâu (deep learning) và các thuật toán học máy (machine learning) để trích xuất thông tin hữu ích từ dữ liệu văn bản [5].

Một trong những ứng dụng chính của AI trong phân tích bình luận khách hàng là phân loại cảm xúc (sentiment classification) và đánh giá mức độ hài lòng dựa trên nội dung đánh giá của người dùng. Các mô hình học sâu, đặc biệt là mạng nơ-ron hồi tiếp (RNN) và bộ nhớ dài ngắn hạn (LSTM), đã chứng minh hiệu quả vượt trội trong việc xử lý ngôn ngữ tự nhiên, giúp cải thiện độ chính xác của hệ thống phân loại bình luận [13]. Ngoài ra, các mô hình transformer, điển hình là BERT (Bidirectional Encoder Representations from Transformers), đã mở ra một hướng tiếp cận mới với khả năng nắm bắt ngữ cảnh từ cả hai chiều của câu, từ đó tăng cường khả năng hiểu và đánh giá nội dung bình luận.

Trong lĩnh vực thương mại điện tử, đặc biệt là trên các nền tảng lớn như Amazon, AI không chỉ giúp tự động hóa quá trình phân loại bình luận mà còn hỗ trợ doanh nghiệp trong việc phát hiện đánh giá giả mạo, đề xuất sản phẩm phù hợp và nâng cao trải nghiệm người dùng. Nghiên cứu ứng dụng AI trong việc phân loại và đánh giá bình luận khách hàng trên các nền tảng thương mại nói chung và trên Amazon nói riêng có ý nghĩa quan trọng trong việc giúp người mua đưa ra quyết định chính xác hơn và hỗ trợ các nhà kinh doanh tối ưu chiến lược tiếp thị.

2.4. Học máy (Machine Learning)

Học máy (Machine Learning - ML) là một lĩnh vực thuộc trí tuệ nhân tạo (AI) với mục tiêu là tạo ra các chương trình máy tính có khả năng tự động cải thiện thông qua kinh nghiệm và sử dụng dữ liệu. Học máy là lĩnh vực nghiên cứu cung cấp cho máy tính khả năng học, thực hiện các công việc thay vì được lập trình một cách rõ ràng [20]. Tuy nhiên, học máy vẫn đòi hỏi sự tham gia của con người trong việc chuẩn bị dữ liệu và tinh chỉnh mô hình để đạt hiệu suất cao. Học máy được phân thành ba phương pháp chính: học có giám sát, học không giám sát, học tăng cường và các phương pháp khác, với mỗi nhiệm vụ và đặc điểm của dữ liệu mà đưa ra lựa chọn phương pháp, mô hình phù hợp nhất [14].

Với sự phát triển của dữ liệu lớn (Big Data) hiện nay thì học máy đã trở thành công cụ quan trọng trong nhiều lĩnh vực như thương mại điện tử, giáo dục, y tế,... Theo Forrester năm 2023, những tiến bộ trong thuật

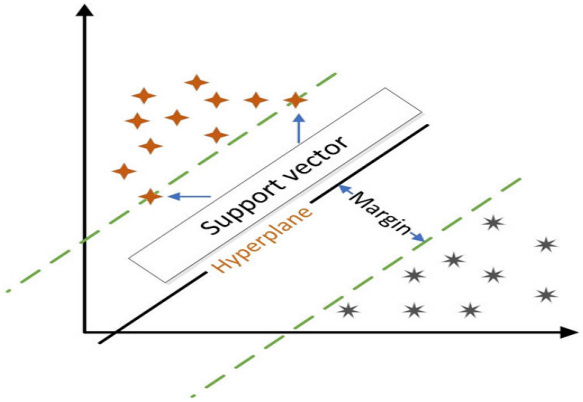
toán học máy mang lại độ chính xác cao và chiều sâu cho phân tích dữ liệu tiếp thị giúp các nhà tiếp thị hiểu các chi tiết tiếp thị. Các mô hình học máy tiên tiến đã được triển khai để phân tích dữ liệu văn bản, đặc biệt trong khai thác ý kiến và phân loại cảm xúc từ đánh giá sản phẩm trên nền tảng thương mại điện tử [1]. Với khả năng xử lý dữ liệu phi cấu trúc tập dữ liệu lớn, học máy đóng vai trò quan trọng trong việc mở ra cơ hội phát triển cho doanh nghiệp bằng cách nâng cao năng suất và tạo ra lợi thế cạnh tranh.

2.5. SVM (Support Vector Machine)

Máy vector hỗ trợ (Support Vector Machine - SVM) là một thuật toán học có giám sát phổ biến trong lĩnh vực trí tuệ nhân tạo và học máy, đặc biệt hiệu quả trong các bài toán phân loại và hồi quy [7]. SVM hoạt động bằng cách tìm một siêu phẳng tối ưu để phân tách các nhóm dữ liệu, đồng thời tối đa hóa khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất (vector hỗ trợ), giúp mô hình tổng quát hóa tốt trên tập dữ liệu mới.

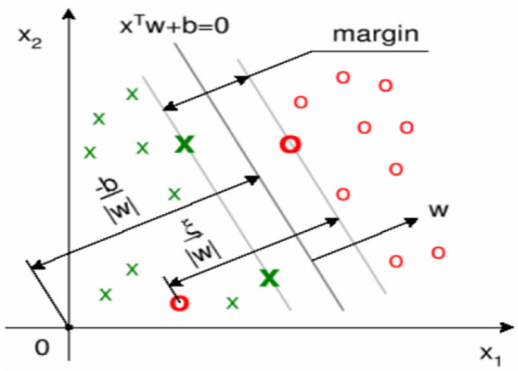
Một mô hình SVM bao gồm: siêu phẳng (Hyperplane) là một đường (trong không gian hai chiều), một mặt phẳng (trong không gian ba chiều) hoặc một mặt phẳng nhiều chiều dùng để phân tách dữ liệu thành các lớp khác nhau. Lề tối ưu (Margin) là khoảng cách từ siêu phẳng phân tách đến các điểm dữ liệu gần nhất của mỗi lớp. Véc-tơ hỗ trợ (Support Vectors) là những điểm dữ liệu nằm gần nhất với siêu phẳng phân tách, đóng vai trò quan trọng trong việc xác định biên của lớp dữ liệu. Hàm mất mát (Loss Function), được xây dựng trên nguyên lý Hinge Loss, đóng vai trò quan trọng trong việc tối ưu hóa mô hình, không chỉ đảm bảo độ chính xác trong phân loại mà còn duy trì khoảng cách tối ưu giữa siêu phẳng quyết định và các điểm dữ liệu. Việc duy trì biên an toàn này góp phần cải thiện khả năng tổng quát hóa của mô hình, hạn chế tình trạng overfitting và nâng cao hiệu suất phân loại trên các tập dữ liệu chưa thấy trước. Tham số C (Regularization Parameter) điều chỉnh sự đánh đổi giữa độ chính xác trên tập huấn luyện và khả năng tổng quát hóa của mô hình. Hàm hạt nhân (Kernel Function) là công cụ giúp SVM có thể làm việc với dữ liệu phi tuyến tính bằng cách ánh xạ dữ liệu vào không gian có số chiều cao hơn. Các hàm hạt nhân điển hình bao gồm: Linear Kernel, Radial Basis Function, Polynomial Kernel, Sigmoid Kernel [10]. Mỗi hàm hạt nhân này có những đặc điểm riêng biệt và phụ thuộc vào đặc tính của dữ liệu và tác vụ cụ thể mà lựa chọn hàm sao cho phù hợp nhất nhằm tối đa hóa kết quả.

Nguyên tắc cốt lõi của SVM là tìm một siêu phẳng tối ưu để phân tách dữ liệu thành các nhóm khác nhau, sao cho khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất của mỗi nhóm (được gọi là margin) là lớn nhất. Việc tối ưu hóa khoảng cách này giúp mô hình có khả năng phân loại chính xác và đảm bảo tính tổng quát cao ngay cả khi dữ liệu có số chiều lớn nhưng số lượng mẫu huấn luyện hạn chế [3]. SVM được sử dụng trong nhiều tác vụ và nhiều lĩnh vực như phân loại văn bản, nhận diện hình ảnh, y học và xe tự lái…



Hình 1. Biểu diễn đồ họa của mô hình SVM *(Nguồn: Aghaabbasi và cộng sự [2])*

Đối với bài toán xử lý văn bản, dữ liệu đầu vào của SVM là chuỗi ký tự nên cần được chuyển đổi thành dạng số. Hai phương pháp phổ biến để thực hiện bước này gồm: TF-IDF (Term Frequency - Inverse Document Frequency): Biểu diễn văn bản bằng cách tính toán tần suất xuất hiện của từ trong tài liệu và điều chỉnh trọng số theo mức độ quan trọng của từ trong tập dữ liệu. Word Embeddings (nhúng từ): Các mô hình như Word2Vec hoặc BERT giúp chuyển đổi từ ngữ thành các vector có nghĩa ngữ nghĩa, nhờ đó giúp các mô hình phân loại như SVM xử lý văn bản một cách hiệu quả hơn, tận dụng các mối quan hệ ngữ nghĩa giữa các từ [21].



Hình 2. Cách hoạt động SVM trong bài toán phân loại tuyến tính *(Nguồn: Moro và cộng sự [15])*

Đánh giá mô hình là một bước quan trọng để kiểm tra hiệu suất và độ tin cậy của mô hình trong bài toán phân loại. Các bình luận mới sẽ được phân loại dựa trên mô hình SVM đã huấn luyện, sau đó đánh giá hiệu suất bằng các chỉ số như độ chính xác (Accuracy), độ nhạy (Recall) và độ đặc hiệu (Precision) [24]. Accuracy là tỷ lệ giữa kết quả dự đoán với dữ liệu thực tế, được dùng để độ chính xác trung bình các thuật toán. Precision cho thấy số lượng dự đoán được thực hiện chính xác hoặc có liên quan trong số tất cả các dự đoán dựa trên lớp tích cực. Recall là chỉ số thể hiện trong tất cả các trường hợp Positive, bao nhiêu trường hợp đã được dự đoán chính xác. Ngoài ra, chỉ số F1-score đo giá trị trung bình hài hòa của độ chính xác (Precision) và độ nhạy (Recall), giúp tối ưu hóa một bộ phân loại cho độ chính xác cân bằng và hiệu suất thu hồi.

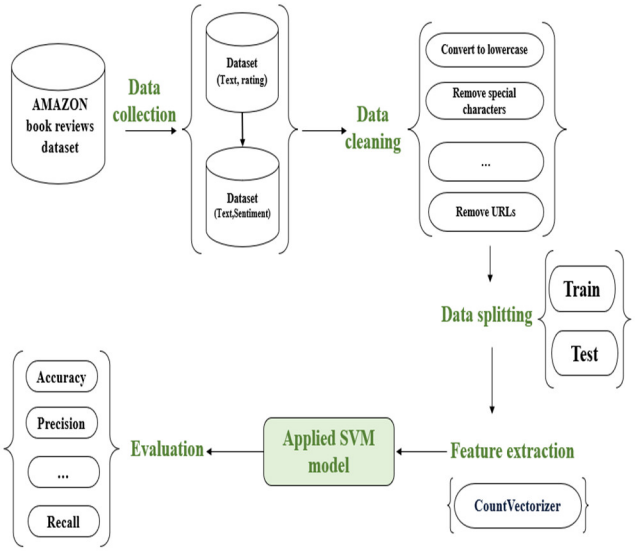
Nhìn chung, SVM là là một công cụ mạnh mẽ trong việc giải quyết cá tác vụ liên quan đến phân loại và sẽ tiếp tục đóng vai trò quan trọng trong các nghiên cứu và ứng dụng công nghệ trong tương lai.

3. PHƯƠNG PHÁP VÀ KẾT QUẢ NGHIÊN CỨU

3.1. Phương pháp nghiên cứu

Bộ dữ liệu được nhóm tác giả sử dụng trong thí nghiệm này là tập dữ liệu đánh giá được thu thập từ trang web https://amazon-reviews-2023.github.io/. Bộ dữ liệu thu thập được là một file nén với số lượng thông tin khổng lồ bao gồm tỉ lệ đánh giá sao, nội dung bình luận, và thông tin về sản phẩm được thu thập trong thời gian từ quý I năm 1996 đến quý III năm 2023. Tuy nhiên, để đạt được hiệu quả phân tích tối đa mà không tốn quá nhiều tài nguyên tính toán, nhóm tác giả đã chọn ngẫu nhiên 100.000 dòng đánh giá ở hai trường nội dung bình luận và đánh giá sao. Trong bài báo này, nhóm tác giả lựa chọn phân tích hai đặc trưng chính của dữ liệu là số sao đánh giá (rating) và nội dung bình luận (text). Đây là những biến thể hiện trực tiếp cảm nhận của người tiêu dùng tại thời điểm đánh giá và nhìn chung ít chịu tác động từ các yếu tố mang tính thời vụ hoặc xu hướng thị trường. Bên cạnh đó, do tập dữ liệu ban đầu không được sắp xếp theo trình tự thời gian, việc lấy mẫu ngẫu nhiên 100.000 dòng từ 1.000.000 dòng đầu tiên được xem là hợp lý, đảm bảo tính khách quan và đại diện cho toàn bộ tập dữ liệu.

Trong đó, các đánh giá trong bộ dữ liệu được chia thành năm cấp độ từ một đến năm sao. Trong các công cụ hỗ trợ phân loại đánh giá bình luận với mô hình SVM, nhóm tác giả lựa chọn Google Colab- một nền tảng do Google phát triển. Google Colab được tích hợp tài nguyên tính toán mạnh mẽ bao gồm GPU (Graphics Processing Unit) và TPU (Tensor Processing Unit) cho phép người dùng thực hiện các tác vụ phân tích phức tạp với tốc độ quá trình huấn luyện và tính toán được tối ưu. Ngoài ra, Google Colab hỗ trợ hầu hết các thư viện Python phổ biến giúp người dùng có dễ dàng thể phát triển các mô hình học sâu.



Hình 3. Quy trình triển khai mô hình SVM *(Nguồn: Tổng hợp kết quả từ nhóm tác giả)*

Quy trình thực hiện mô hình SVM trong phân tích dữ liệu đánh giá bao gồm các bước chính: Thu thập dữ liệu, xử lý dữ liệu, trích xuất đặc trưng, huấn luyện mô hình và đánh giá hiệu suất.

Dữ liệu đánh giá, bao gồm nội dung bình luận và số sao, được thu thập nhằm phục vụ phân tích cảm xúc. Nhóm nghiên cứu tiến hành gán nhãn thủ công dựa trên thang điểm đánh giá, trong đó các bình luận có số sao từ 1 đến 3 được phân loại là tiêu cực (negative), còn các đánh giá từ 4 đến 5 sao được xem là tích cực (positive). Việc gán nhãn này giúp tạo ra một tập dữ liệu có nhãn rõ ràng, hỗ trợ huấn luyện mô hình học máy để tự động phân loại cảm xúc từ các bình luận mới. Tiếp theo, dữ liệu được tiền xử lý nhằm nâng cao chất lượng đầu vào, bao gồm loại bỏ ký tự đặc biệt, chuyển đổi văn bản về dạng chữ thường và thực hiện các bước chuẩn hóa khác nhằm giảm thiểu sai sót khi huấn luyện mô hình. Sau khi xử lý, tập dữ liệu được chia thành hai phần: 70% dữ liệu được sử dụng để huấn luyện mô hình (train data), trong khi 30% còn lại được dùng để kiểm tra và đánh giá hiệu suất của mô hình (test data). Phương pháp này không chỉ cải thiện độ chính xác của mô hình mà còn hỗ trợ các nghiên cứu về phản hồi của người dùng trong nhiều lĩnh vực khác nhau.

Trong giai đoạn tiếp theo, quá trình trích xuất đặc trưng từ văn bản được nhóm nghiên cứu thực hiện nhằm chuyển đổi dữ liệu văn bản thành dạng biểu diễn số học, tạo điều kiện cho mô hình phân loại cảm xúc. Nhóm nghiên cứu áp dụng phương pháp Bag-of-Words (BoW), trong đó mỗi bình luận được biểu diễn dưới dạng một vectơ đặc trưng phản ánh tần suất xuất hiện của từng từ trong tập từ vựng. Cách tiếp cận này giúp mô hình Support Vector Machine (SVM) có thể tiếp nhận và xử lý dữ liệu hiệu quả hơn. Sau khi hoàn tất quá trình trích xuất đặc trưng, mô hình SVM được huấn luyện trên tập dữ liệu huấn luyện (training set) nhằm tìm kiếm siêu phẳng tối ưu (optimal hyperplane) để phân tách hai nhóm cảm xúc: tích cực và tiêu cực. Tiếp đó, mô hình được kiểm thử trên tập dữ liệu kiểm tra (test set) để đánh giá khả năng tổng quát hóa. Hiệu suất của mô hình được đo lường thông qua các chỉ số đánh giá như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (precision) và F1-score. Từ đó cung cấp cơ sở để đánh giá mức độ hiệu quả của mô hình so với kỳ vọng ban đầu. Kết quả thu được không chỉ giúp xác định khả năng ứng dụng thực tế của mô hình mà còn làm cơ sở để đề xuất các hướng cải tiến nhằm nâng cao chất lượng phân loại trong các nghiên cứu tiếp theo.

3.2. Kết quả nghiên cứu

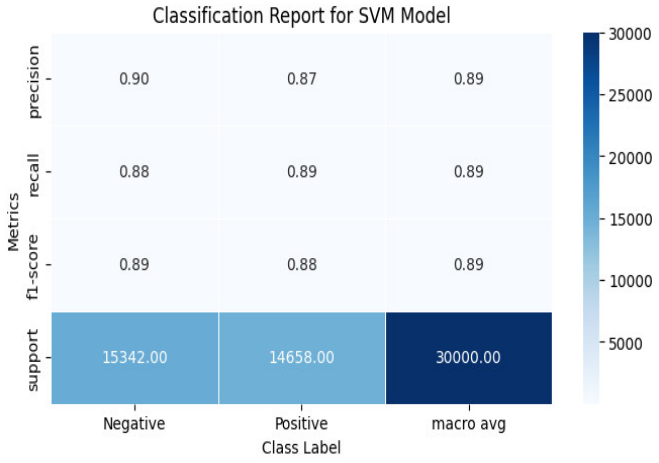
Bảng 1. Bảng phân tích kết quả phân loại bình luận

Bình luận	Phân tích	Kết quả phân loại
"I love this book and so does my emotional son. It is definitely helpful, and it is so relaxing."	Văn bản được chuyển đổi thành vector BoW, với các từ như *"love"*, *"helpful"*, *"relaxing"* xuất hiện với tần suất cao. Những từ này thường liên quan đến cảm xúc tích cực, giúp mô hình xác định hướng phân loại.	"Positive". Tổng trọng số của các từ trong văn bản này nghiêng về lớp tích cực, do đó SVM phân loại là đánh giá tích cực.
"This has been a great series with the characters evolving in each new book. It was a page-turner and I read it on a snowy winter day!"	Các từ như *"great"*, *"evolving"*, *"page-turner"* xuất hiện với tần suất cao và thường liên quan đến các đánh giá tích cực trong tập huấn luyện. SVM sử dụng các trọng số đã học để phân loại văn bản này.	"Positive". Do có nhiều từ mang sắc thái tích cực, tổng trọng số của vectơ này nghiêng về phía lớp tích cực theo siêu phẳng tối ưu của SVM.
"I felt like I was reading a novel where the author got up every morning and put down a few paragraphs, not remembering what he wrote the day before."	Các từ như *"not remembering"*, *"few paragraphs"* thường xuất hiện trong các đánh giá tiêu cực, làm cho vector đặc trưng của văn bản này nghiêng về phía lớp tiêu cực.	"Negative". Mô hình SVM xác định rằng tổng trọng số của các từ trong văn bản này nằm phía lớp tiêu cực của siêu phẳng quyết định.

(Nguồn: Tổng hợp kết quả từ nhóm tác giả)

Trong nghiên cứu này, bằng phương pháp trích xuất Bag of Words (BoW) thì với những từ có tính chất biểu cảm mạnh, chẳng hạn như *"waste"*, *"boring"*, *"disappointed"*, có thể giúp mô hình nhận diện các đánh giá tiêu cực, trong khi các từ như *"love"*, *"amazing"*, *"perfect"* thường xuất hiện trong những đánh giá tích cực. Để đánh giá hiệu quả của mô hình, nhóm tác giả tiến hành phân tích một số bình luận mẫu.

Để đánh giá hiệu suất của mô hình, các chỉ số như precision, recall, F1-score và accuracy được phân tích. Dưới đây là kết quả phân loại và các giá trị đo lường hiệu suất, phản ánh độ chính xác cũng như khả năng nhận diện cảm xúc của mô hình (hình 4 và 5).



Hình 4. Báo cáo phân loại trực quan của mô hình SVM *(Nguồn: Tổng hợp kết quả từ nhóm tác giả)*

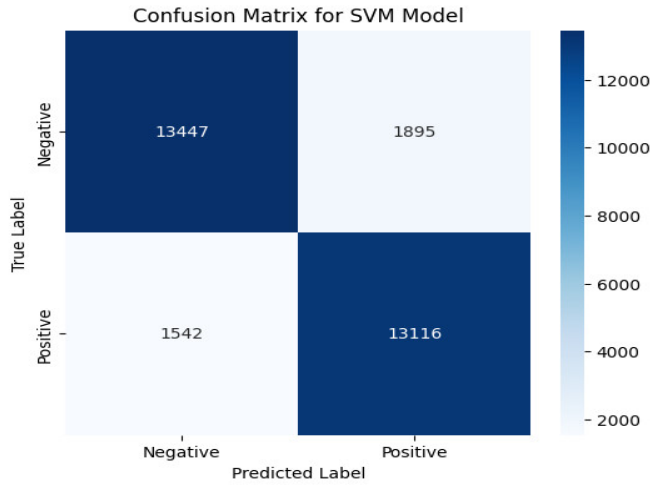
	precision	recall	f1-score	support
Negative	0.90	0.88	0.89	15342
Positive	0.87	0.89	0.88	14658
accuracy			0.89	30000
macro avg	0.89	0.89	0.89	30000
weighted avg	0.89	0.89	0.89	30000

Hình 5. Báo cáo hiệu suất phân loại chi tiết của mô hình SVM *(Nguồn: Tổng hợp kết quả từ nhóm tác giả)*

Thông qua hình 4 và 5 có thể thấy, với việc sử dụng mô hình SVM trong phân tích đánh giá khách hàng trên

sàn thương mại điện tử Amazon cho thấy mô hình đạt độ chính xác (accuracy) cao, lên đến 89%. Xét về từng lớp, mô hình đạt precision là 0,90 đối với lớp Negative và 0,87 đối với lớp Positive. Điều này cho thấy khi mô hình dự đoán đánh giá tiêu cực tốt hơn so với khi dự đoán một đánh giá là tích cực. Tuy nhiên, độ nhạy (recall) của hai lớp gần như cân bằng, lần lượt là 0,88 và 0,89, chứng tỏ mô hình có khả năng phát hiện tốt cả hai loại đánh giá mà không bỏ sót đáng kể. Giá trị F1-score, thước đo tổng hợp giữa precision và recall, đạt 0,89 cho lớp Negative và 0,88 cho lớp Positive, phản ánh sự đồng đều trong hiệu suất phân loại. Số lượng mẫu giữa hai lớp khá cân bằng (15.342 mẫu Negative và 14.658 mẫu Positive), giúp mô hình duy trì tính ổn định. Bên cạnh đó, cả hai giá trị trung bình macro và trung bình có trọng số (weighted avg) đều đạt 0,89, cho thấy sự ổn định và cân bằng của mô hình trên tập dữ liệu thử nghiệm.

Để đánh giá một cách toàn diện hiệu suất mô hình phân loại đánh giá dựa trên SVM, nhóm tác giả tiến hành trực quan hóa ma trận nhầm lẫn. Ma trận này cung cấp một cái nhìn tổng quan và chi tiết về khả năng phân loại của mô hình. Các dự đoán đúng và không chính xác được đánh dấu và chia thành các lớp. Kết quả của các dự đoán được so sánh với các giá trị thực. Bằng cách phân tích ma trận nhầm lẫn, có thể xác định được những bình luận như thế nào thì có khả năng bị phân loại sai cao nhất, từ đó đưa ra các biện pháp can thiệp phù hợp để nâng cao độ tin cậy của mô hình dự báo.



Hình 6. Ma trận nhầm lẫn *(Nguồn: Tổng hợp kết quả từ nhóm tác giả)*

Trong tổng số 15.342 đánh giá thực tế thuộc lớp Negative, mô hình phân loại chính xác 13.447 trường hợp, nhưng vẫn nhầm lẫn 1.895 trường hợp sang lớp Positive (tương đương với 12,35% bị nhầm lẫn). Tương tự, trong 14.658 đánh giá thực tế thuộc lớp Positive, mô hình dự đoán đúng 13.116 trường hợp, trong khi 1.542 đánh giá bị nhầm sang lớp Negative (khoảng 10,52%). Dù tỷ lệ nhầm lẫn không phải là quá thấp, nhưng vẫn có thể thấy số lượng dự đoán đúng chiếm ưu thế rõ rệt so với số lượng sai. Tỷ lệ nhầm lẫn ở mức khoảng 10 - 12% được xem là khá tốt với bài toán phân loại cảm xúc được đặt ra, đặc biệt khi xét đến sự phức tạp của ngôn ngữ tự nhiên, điều này gây khó khăn cho mô hình SVM trong việc phân biệt giữa cảm xúc tích cực và tiêu cực, do ngữ cảnh và sắc thái cảm xúc trong văn bản có thể dẫn đến sai sót trong quá trình phân loại.

Kết quả nghiên cứu cho thấy mô hình SVM có khả năng phân loại hiệu quả, giúp tự động hóa quá trình phân tích phản hồi của khách hàng một cách nhanh chóng và chính xác. Mô hình SVM không chỉ cho kết quả ổn định mà còn có thể xử lý tốt các dữ liệu lớn, giúp tiết kiệm thời gian và công sức trong việc phân tích lượng lớn thông tin từ các đánh giá khách hàng. Ngoài ra, SVM triển khai đơn giản, dễ dàng áp dụng vào các hệ thống phân tích mà không đòi hỏi quá nhiều tài nguyên tính toán. Tuy nhiên, vẫn còn tồn tại một số hạn chế, đặc biệt là các trường hợp bị phân loại sai do sắc thái trung lập hoặc câu văn phức tạp. Để khắc phục điều này, có thể cân nhắc tối ưu hóa đặc trưng đầu vào, áp dụng kỹ thuật giảm nhiễu, hoặc kết hợp với các mô hình ngôn ngữ tiên tiến để cải thiện khả năng hiểu ngữ nghĩa.

Nhìn chung, SVM là một phương pháp mạnh mẽ, ổn định và có tiềm năng ứng dụng thực tiễn trong phân tích đánh giá khách hàng, góp phần hỗ trợ các nền tảng thương mại điện tử trong việc cải thiện dịch vụ và đưa ra quyết định dựa trên ý kiến người dùng.

4. KẾT LUẬN

Nghiên cứu đã khẳng định hiệu quả của mô hình SVM (Support Vector Machine) trong việc phân loại cảm xúc từ các bình luận của khách hàng trên nền tảng thương mại điện tử Amazon. SVM đã chứng minh được khả năng xử lý dữ liệu văn bản một cách hiệu quả, đặc biệt trong các tác vụ phân loại đa lớp nhờ vào tính ổn định cao và khả năng tổng quát hóa tốt. Bên cạnh đó, bằng cách sử dụng siêu phẳng tối ưu để phân tách dữ liệu mà mô hình SVM có thể nhận diện chính xác các đánh giá tích cực, tiêu cực từ khách hàng. Từ đó, hỗ trợ doanh nghiệp trong việc khai thác thông tin quan trọng về ý kiến người tiêu dùng.

Từ kết quả nghiên cứu có thể thấy, SVM được áp dụng rộng rãi để giúp doanh nghiệp theo dõi xu hướng cảm xúc của khách hàng làm tiền đề giúp doanh nghiệp đưa ra các chiến lược tiếp thị, cải thiện chất lượng dịch

vụ và tối ưu hóa sản phẩm. Bên cạnh đó, SVM nổi trội hơn về tính hiệu quả với tập dữ liệu vừa và nhỏ, yêu cầu tài nguyên tính toán thấp hơn nhưng vẫn đạt độ chính xác cao khi được kết hợp với các kỹ thuật tiền xử lý văn bản phù hợp. Điều này giúp các doanh nghiệp, đặc biệt là những công ty vừa và nhỏ, có thể triển khai phân tích cảm xúc mà không cần đầu tư kinh phí lớn vào hạ tầng công nghệ. Việc sử dụng SVM để phân loại đánh giá của khách hàng không chỉ là việc phân loại cảm xúc, mà còn có thể hỗ trợ xây dựng các hệ thống phản hồi tự động, giúp doanh nghiệp nhanh chóng phát hiện các vấn đề trong sản phẩm, dịch vụ và có những điều chỉnh kịp thời. Hơn nữa, mô hình này có thể được tích hợp vào các công cụ quản lý quan hệ khách hàng (CRM), giúp doanh nghiệp cá nhân hóa chiến lược tiếp cận khách hàng dựa trên cảm xúc và nhu cầu thực tế của họ.

Ngoài những ưu điểm trên thì mô hình Support Vector Machine (SVM) vẫn có một số hạn chế. Trước tiên, SVM phụ thuộc đáng kể vào các kỹ thuật trích xuất đặc trưng và tiền xử lý dữ liệu, yêu cầu quá trình thiết lập và tối ưu hóa cẩn trọng để đạt hiệu suất phân loại cao. Đặc biệt, việc lựa chọn hàm kernel phù hợp có tác động đáng kể đến độ chính xác của mô hình, đòi hỏi quá trình điều chỉnh tham số nhiều lần để tối ưu hóa hiệu suất. Ngoài ra, SVM gặp khó khăn trong việc nắm bắt ngữ nghĩa ngữ cảnh và mối quan hệ giữa các từ, đặc biệt khi áp dụng trên các tập dữ liệu lớn và phức tạp. Những hạn chế này đặt ra hướng nghiên cứu trong tương lai nhằm cải thiện khả năng của SVM trong phân tích cảm xúc. Một trong những giải pháp tiềm năng là kết hợp SVM với các mô hình biểu diễn từ tiên tiến như Word2Vec, TF-IDF hoặc BERT embeddings để nâng cao khả năng học đặc trưng ngữ nghĩa. Bên cạnh đó, các nghiên cứu tiếp theo có thể tập trung vào việc tối ưu hóa thuật toán và điều chỉnh siêu tham số nhằm tăng cường khả năng tổng quát hóa của mô hình. Việc tiếp tục cải tiến SVM trong phân tích cảm xúc không chỉ giúp nâng cao độ chính xác của mô hình mà còn hỗ trợ doanh nghiệp khai thác dữ liệu khách hàng hiệu quả hơn, góp phần nâng cao năng lực cạnh tranh trong bối cảnh thương mại điện tử ngày càng phát triển.

TÀI LIỆU THAM KHẢO

[1]. Agarwal B., Mittal N., Agarwal B., Mittal N., "Machine learning approach for sentiment analysis," *Prominent feature extraction for sentiment analysis*, 21-45, 2016.

[2]. Aghaabbasi M., Ali M., Jasiński M., Leonowicz Z., Novák T., "On hyperparameter optimization of machine learning methods using a Bayesian optimization algorithm to predict work travel mode choice," *IEEE Access, 11, 19762-19774, 2023*.

[3]. Burges C. J., "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 2(2), 121-167, 1998.

[4]. Cambria E., Schuller B., Xia Y., Havasi C., "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems, 28(2),*15-21, 2013.

[5]. Cambria E., Song Y., Wang H., Howard N., "Semantic multidimensional scaling for open-domain sentiment analysis," *IEEE intelligent systems*, *29*(2), 44-51, 2012.

[6]. Clissa L., Lassnig M., Rinaldi L., "How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry," *Frontiers in big Data*, 6, 1271639, 2023.

[7]. Cortes C., Vapnik V., "Support-vector networks," *Machine Learning*, 20(3), 273-297, 1995

[8]. Chen M., Mao S., Liu Y., "Big data: A survey," *Mobile networks and applications*, 19, 171-209, 2014.

[9]. Gandomi A., Haider M., "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, 35(2), 137-144, 2015.

[10]. Howley T., Madden M. G., "The genetic kernel support vector machine: Description and evaluation," *Artificial Intelligence Review*, 24,379-395, 2005.

[11]. Kumar V., Reinartz W., *Customer Relationship Management*. Springer-Verlag GmbH Germany, part of Springer Nature 2006, 2012, 2018.

[12]. Laney D., "3D data management: Controlling data volume, velocity, and variety," *META group research note*, 6(70), 1, 2001.

[13]. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V., "Roberta: A robustly optimized bert pretraining approach," 2019. *arXiv preprint arXiv*:1907.11692.

[14]. Mahesh B., "Machine learning algorithms - a review," *International Journal of Science and Research (IJSR),* 9(1), 381-386, 2020.

[15]. Moro R. A., Härdle W., Schäfer D., *Rating companies with support vector machines*. Germany: German Institute for Economic Research, 2004.

[16]. Pang B., Lee L., "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, 2(1–2), 1–13, 2008.

[17]. Plutchik R., "A general psychoevolutionary theory of emotion," in R. Plutchik, *Theories of emotion,* 3-33, 1980.

[18]. Russell S. J., Norvig P., *Artificial intelligence: a modern approach*. Pearson, 2016.

[19]. Samuel I., Ogunkeye F. A., Olajube A., Awelewa A., "Development of a voice chatbot for payment using Amazon Lex service with eyowo as the payment platform," *in 2020 International Conference on Decision Aid Sciences and Application (DASA)*, IEEE, 104-108, 2020.

[20]. Samuel A., "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, 3(3), 210-229, 1959.

[21]. Sebastiani F., "Machine learning in automated text categorization," *ACM Computing Surveys*, 34(1), 1-47, 2002.

[22]. Sharma R., "Amazon Dreams of AI Agents That Do the Shopping for You," *Wired*, 2023.

[23]. Snijders C. C. P., Matzat U., Reips U. D., ""Big Data": big gaps of knowledge in the field of internet science," *International journal of internet science*, 7(1), 1-5, 2012.

[24]. Yang Y., Liu X., "A re-examination of text categorization methods," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 42-49, Association for Computing Machinery, New York 1999.

[25]. Zhang X., Zhao J., LeCun Y., "Character-level convolutional networks for text classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems,* 649–657, MIT Press, 2015.

AUTHORS INFORMATION

Nguyen Thi Hong Duyen[1], Vu Thi Hong Anh[2]

[1]School of Economics, Hanoi University of Industry, Vietnam

[2]Students majoring in Business Data Analytics, Hanoi University of Industry, Vietnam