PHÂN TÍCH Ý KIẾN KHÁCH HÀNG TRỰC TUYẾN DỰA THEO PHƯƠNG PHÁP HỌC MÁY

Bùi Minh Hiển
Đại học Kinh tế Thành phố Hồ Chí Minh
Email: hienbui.192118003@st.ueh.edu.vn
Nguyễn Thành Phát
Đại học Kinh tế Thành phố Hồ Chí Minh
Email: phatnguyen.192118008@st.ueh.edu.vn
Phạm Thị Thiên Hương
Đại học Kinh tế Thành phố Hồ Chí Minh
Email: huongpham.192118006@st.ueh.edu.vn
Nguyễn Thị Bảo Hương
Đại học Kinh tế Thành phố Hồ Chí Minh
Email: huongnguyen.192118005@st.ueh.edu.vn
Hồ Trung Thành
Đại học Kinh tế - Luật, Đại học Quốc gia Hồ Chí Minh
Email: thanhht@uel.edu.vn

Mã bài: JED - 131020
Ngày nhận: 13/10/2020
Ngày nhận bản sửa: 29/12/2020
Ngày duyệt đăng: 05/9/2021

Tóm tắt:
Phân tích cảm xúc hay khai phá ý kiến dựa trên những phản hồi của khách hàng trước, trong và sau mua sắm đóng vai trò rất quan trọng để doanh nghiệp xây dựng chiến lược kinh doanh phù hợp đối với từng sản phẩm, dịch vụ hay đối với từng phân khúc khách hàng. Thông qua việc khảo sát các mô hình phân tích và hiểu ý kiến khách hàng, bài báo trước hết tập trung vào đề xuất mô hình phân tích ý kiến khách hàng trực tuyến và thử nghiệm phương pháp với trường hợp cụ thể là tập dữ liệu được thu thập từ ứng dụng thương mại điện tử Lazada – một trong các sàn thương mại điện tử hàng đầu tại Việt Nam với nhiều năm đứng đầu thị trường. Tiếp theo, nhóm tác giả dựa vào phương pháp học máy có giám sát với hai thuật toán hồi quy Logistic và Random Forest để thực nghiệm mô hình, so sánh và đánh giá độ chính xác. Kết quả nghiên cứu hàm ý phương pháp phân tích và thấu hiểu trải nghiệm khách cho nhà quản lý để từ đó triển khai có cơ sở xây dựng chiến lược kinh doanh phù hợp hơn.
Từ khóa: Phân tích ý kiến khách hàng, thương mại điện tử, khách hàng trực tuyến, phân tích cảm xúc, học máy có giám sát.
Mã JEL: C61, C67, M00, M3.

Analyzing online customers' reviews based-on machine learning methods

Abstract:
Sentiment analysis and opinion mining based on customers' reviews before, during, and after shopping are very important for businesses to build a suitable business strategy for each product, service, and online customer segment. Through surveying sentiment analysis models and understanding customers' opinions, this article firstly is to propose the model of online customers' opinion analysis and to test method with a dataset which is collected from the e-commerce application of Lazada company - one of the leading e-commerce site in Vietnam's market in many years. Then, the supervised machine learning methods with Logistic Regression and Random Forest are applied to experiment with the proposed model, compare and evaluate accuracy as well. The research results recommend the method to analyze and to understand customer experience to develop a more suitable business strategy.
Keywords: *Analyzing reviews of customers, e-commerce, online customers, sentiment analysis, supervised machine learning.*
JEL codes: C61, C67, M00, M3.

1. Giới thiệu

Kỷ nguyên của thông tin điện tử trong mọi giai đoạn của cuộc sống phát triển nhanh chóng và sản sinh ra một số lượng lớn dữ liệu người dùng đang hoạt động với quy mô các bài đánh giá của họ được tạo ra hàng ngày trực tuyến trên các trang web lớn. Hệ thống phân tích tự động nghĩa là phân tích, tóm tắt, phân loại dữ liệu và các phương pháp hiệu quả để lưu trữ lượng dữ liệu khổng lồ. Khai phá văn bản là một cách tiếp cận được sử dụng vào các lĩnh vực khác nhau như máy học, truy xuất thông tin, thống kê và ngôn ngữ học tính toán để khai phá ý kiến. Mục tiêu của phân tích cảm xúc là làm ra máy học tự động có khả năng nhận dạng và phân loại cảm xúc. Suy nghĩ, quan điểm và thái độ dựa trên biểu cảm thay vì lý trí được gọi là cảm xúc (Kaushik & cộng sự, 2015).

Phân tích cảm xúc, khai phá ý kiến liên quan đến đối tượng hoặc chủ đề được thảo luận. Đối mặt với khó khăn của việc đánh giá, chọn lọc cảm xúc phù hợp là phải hiểu một phần lời nói của ngôn ngữ vì ý nghĩa chính xác là một điều rất quan trọng trong việc đưa ra quyết định, bên cạnh đó phải đảm bảo tiết kiệm về mặt thời gian và công sức khi nghiên cứu. Quá trình này đòi hỏi một từ vựng khổng lồ để xử lý cảm xúc và phân biệt đâu là tích cực, đâu là tiêu cực. Một số bình luận bao gồm spam, giả mạo hoặc trùng lặp là một trong những thách thức có ảnh hưởng cũng như gây trở ngại đến sự hiểu biết về nhận xét, đánh giá cảm xúc của người tiêu dùng (Hussein, 2016). Phân tích cảm xúc khách hàng cũng được nghiên cứu và ứng dụng trong nhiều lĩnh vực khác nhau, trong đó, nhóm tác giả Le & cộng sự (2017) đã áp dụng phương pháp học máy mạng nơ-ron trong phân tích cảm xúc từ dữ liệu là ý kiến của khách hàng để lại trên mạng xã hội để dự đoán sự quan tâm của khách hàng trong ngành thức ăn nhanh.

Dữ liệu do người dùng tạo mang nhiều thông tin giá trị về sản phẩm, con người, sự kiện... nói riêng và hành vi khách hàng nói chung. Bước đầu tiên trong phân loại cảm xúc là tiền xử lý văn bản, quá trình này sẽ làm cho dữ liệu phi cấu trúc có chứa nhiễu hiển thị trên web cũng như một hình thức được sử dụng để phân loại. Tiền xử lý bao gồm các nhiệm vụ như làm sạch, tách từ, loại bỏ những từ không có nghĩa, ký tự không cần thiết, chuẩn hóa chữ thường, chữ hoa. Giai đoạn tiếp theo là các trích xuất đặc trưng. Có nhiều loại phương pháp trích xuất đặc trưng như: túi từ (Bag-of-words), tần suất xuất hiện của từ - tần suất nghịch đảo văn bản (Term frequency/Inverse document frequency - TF-IDF) dựa trên phương pháp xử lý ngôn ngữ tự nhiên. Giai đoạn cuối cùng là lựa chọn, xây dựng phương pháp và áp dụng các thuật toán mô hình máy học phổ biến như thuật toán Cây quyết định (Decision tree), Máy vector hỗ trợ (Support Vector Machines - SVM), phân lớp Naïve Bayes, Mạng nơ-ron nhân tạo (Artificial neural network) để phân loại và dùng công cụ để biểu diễn trực quan dữ liệu trên dashboard (Ahuja & cộng sự, 2019). Dựa vào thông tin và tri thức có được từ các phương pháp trên, các nhà quản trị kinh doanh nắm bắt thông tin nhanh chóng và có thêm chiều thông tin dễ dàng hơn ra quyết định chính xác hơn từ đó cải tiến mức độ hài lòng và kỳ vọng của khách hàng. Phần còn lại của bài báo bao gồm:

Phần 2 đưa ra tổng quan nghiên cứu và cơ sở lý thuyết; Phương pháp nghiên cứu chi tiết ở phần 3; Kết quả và thảo luận được trình bày ở phần 4; Cuối cùng, phần 5 là kết luận.

2. Tổng quan nghiên cứu và cơ sở lý thuyết

2.1. Các nghiên cứu liên quan

Thông qua những nghiên cứu bằng phương pháp Mạng nơ-ron hai chiều (Bidirectional recurrent neural networks) của Schuster & Paliwal (1997), phân tích cảm xúc cho các sản phẩm của Amazon bằng cách sử dụng Rừng cách ly của Salmiah & cộng sự (2019), phân tích cảm xúc sử dụng thuật toán Rừng ngẫn nhiên trên truyền thông mạng xã hội của Bahrawi (2019) và phương pháp Naïve Bayes của Mali & cộng sự (2016) về lĩnh vực thương mại điện tử, phân tích văn bản là một lĩnh vực nghiên cứu trong ngôn ngữ học máy tính và xử lý ngôn ngữ tự nhiên mà các công ty đang bắt đầu chuyển sang lắng nghe trên mạng xã hội để hiểu khách hàng của họ, để cải thiện sản phẩm hoặc dịch vụ. Bài báo tập trung vào ứng dụng phương pháp phân loại là hồi quy Logistic và Random Forest và kèm theo phương pháp kiểm chứng chéo k-fold để dự đoán mô hình chính xác hơn trong phân tích ý kiến khách hàng trực tuyến. Và cuối cùng là so sánh độ hiệu quả giữa các phương pháp để chọn ra phương pháp phù hợp với bộ dữ liệu đầu vào.

2.2. Xử lý ngôn ngữ tự nhiên (NLP)

Xử lý ngôn ngữ tự nhiên dựa trên nghiên cứu của Pham (2018) là một lĩnh vực đặc trưng, là sự kết hợp giữa các ngành khoa học máy tính, trí tuệ nhân tạo và ngôn ngữ học. Mục tiêu của việc xử lý ngôn ngữ tự

nhiên là để cho máy tính xử lý và hiểu được ngôn ngữ tự nhiên của con người, giúp máy tính có thể thực hiện được một số nhiệm vụ hữu ích thay cho con người như đặt lịch hẹn, mua bán hàng hóa, dịch từ ngôn ngữ này sang ngôn ngữ khác, các hệ tư vấn, hệ hỏi đáp, chẳng hạn như Apple Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, ...

Xử lý ngôn ngữ tự nhiên thường được chia thành các cấp độ khác nhau. Trong đó, đầu vào thường là hai dạng chính của ngôn ngữ gồm lời nói (speech) và văn bản (text). Sau khi phân tích ngữ âm (dạng speech) hoặc OCR/Tokenization văn bản, phương pháp trải qua các bước xử lý ngôn ngữ theo các cấp độ (Pham, 2018) bao gồm: phân tích hình thái của ngôn ngữ; phân tích cú pháp, tìm hiểu cấu trúc của câu (chủ ngữ, động từ chính...); diễn dịch ngữ nghĩa, ý nghĩa của câu dựa vào các từ tạo nên câu phân tích ngữ nghĩa dựa trên bối cảnh của câu.

2.3. Phương pháp học máy

Theo nghiên cứu của Trần Thị Ngọc Thảo & cộng sự (2014), định nghĩa học máy là một nhánh thuộc trí tuệ nhân tạo, sử dụng kiến thức của xác suất thống kê và đại số tuyến tính để nghiên cứu phát triển các kỹ thuật và xây dựng các chương trình mà máy tính có thể học hỏi, phân tích các dữ liệu đã có và tạo ra kết quả cho các dữ liệu mới. Học máy có rất nhiều ứng dụng thực tế như phân tích ý kiến đánh giá, phân tích thị trường, phát hiện thẻ tín dụng giả, chẩn đoán y khoa, nhận dạng tiếng nói và chữ viết, phân loại email, dịch tự động, cử động robot và còn nhiều ứng dụng khác. Một số dạng thuật toán thường dùng được phân loại dựa trên kết quả đầu ra mong muốn hoặc loại đầu vào trong quá trình huấn luyện máy là: học có giám sát (supervised learning), học không giám sát (unsupervised learning), học bán giám sát (semi-supervised), học tăng cường (reinforcement learning) (Trần Thị Ngọc Thảo & cộng sự, 2014).

Trong bài báo, phương pháp và mô hình đề xuất được thử nghiệm bằng phương pháp học có giám sát với hai thuật toán Hồi quy Logistic (Logistic regression) và Rừng ngẫu nhiên (Random forests) (Ho, 1995). Đây là hai thuật toán được sử dụng để phân loại nhị phân (Binary classification). Nhiệm vụ của các thuật toán sẽ phân loại các thành phần của một tập hợp thành hai nhóm dựa trên cơ sở các tiêu chí để phân loại. Điển hình trong nghiên cứu này, phương pháp học máy trên được sử dụng để phân tích và phân loại phản hồi của các khách hàng trên trang thương mại điện tử thuộc nhóm tiêu cực hoặc tích cực liên quan đến các sản phẩm và dịch vụ.

2.4. Hồi quy logistic (Logistic regression)

Đỗ Minh Hải (2017) cho rằng phương pháp hồi quy logistic là một mô hình hồi quy nhằm dự đoán giá trị đầu ra rời rạc y, ứng với một véc-tơ đầu vào X. Việc này tương đương với phương pháp phân loại các giá trị đầu vào X vào các nhóm y tương ứng.

Hồi quy Logistic là một trong những thuật toán học máy đơn giản và dễ thực hiện nhưng lại mang hiệu quả hơn các phương pháp khác trong nhiều trường hợp. Cũng vì lý do này, việc huấn luyện một mô hình với thuật toán này không đòi hỏi khả năng hệ thống tính toán cao. Thuật toán này cho phép mô hình được cập nhật dễ dàng hơn với dữ liệu mới. Hồi quy Logistic đưa ra mô hình xác suất được hiệu chỉnh tốt cùng với kết quả phân loại. Đây là một lợi thế so với các mô hình chỉ đưa ra kết quả phân loại cuối cùng.

Phương pháp thống kê cho rằng khả năng một đầu vào X nằm vào một nhóm y_0 là xác suất nhóm y_0 khi biết X: $p(y_0|X)$. Dựa vào công thức xác xuất hậu nghiệm ta có:

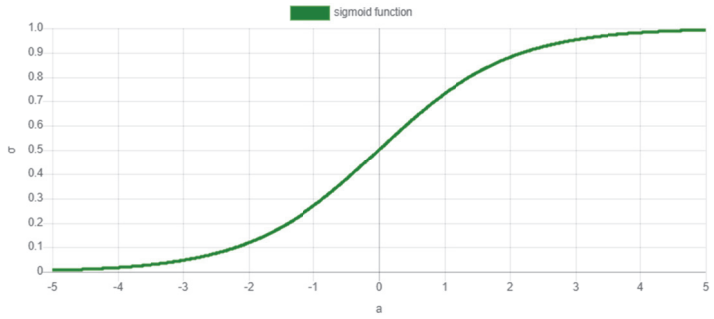
$$p(y_0|x) = \frac{p(x|y_0)p(y_0)}{p(x)} = \frac{p(x|y_0)p(y_0)}{p(x|y_0)p(y_0) + p(x|y_1)p(y_1)} \quad \textbf{(1)}$$

Đặt $$a = ln\frac{p(x|y_0)p(y_0)}{p(x|y_1)p(y_1)} \quad \textbf{(2)}$$

Ta có $$P(y_0|x) = \frac{1}{1 + exp^{-a}} = \sigma(a) \quad \textbf{(3)}$$

Hàm σ(a) ở đây được gọi là hàm sigmoid (*logistic sigmoid function*).

Hình 1: Đồ thị hàm sigmoid σ(a)



Nguồn: Đỗ Minh Hải (2017).

Vận dụng thuyết phân phối chuẩn, ta có thể chỉ ra rằng:

$$a = W^T X + w_0 \quad (4)$$

Đặt: $X_0 = [1,\ldots,1]$, ta có thể viết gọn lại thành:

$$a = W^T X \quad (5)$$

Công thức tính xác suất lúc này:

$$P(y_0|X) = \frac{1}{1+exp^{-a}} = \sigma(W^T X) \quad (6)$$

Trong đó, X là thuộc tính đầu vào; còn W là trọng số tương ứng. Tính được xác suất của công thức sau đó ta có thể sử dụng một ngưỡng $\varepsilon \in [0,1]$ để quyết định nhóm tương ứng. Cụ thể:

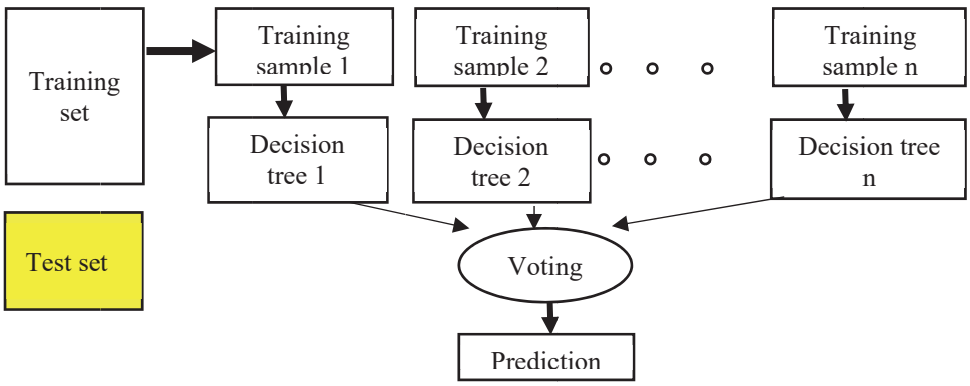
$$\begin{cases} X \in y_0 \text{ if } p(y_0|X) \geq \varepsilon \\ X \in y_1 \text{ if } p(y_1|X) < \varepsilon \end{cases} \quad (7)$$

Ví dụ, $\varepsilon = 0,7$ thì $X \in y_0$ khi mà xác suất thuộc nhóm y_0 của nó là trên 70%, còn dưới 70% thì ta phân nó vào nhóm y_1.

2.5. Random forest

Random forests (RF) được giới thiệu bởi Ho (1995) là phương pháp học có giám sát (Supervised learning). Phương pháp này có thể được sử dụng cho cả phân lớp và hồi quy và cũng là thuật toán linh hoạt và dễ ứng dụng (Hình 2).

Hình 2: Qui trình bốn bước hoạt động của thuật toán Random Forests



Nguồn: Nguyễn Duy Sim (2018).

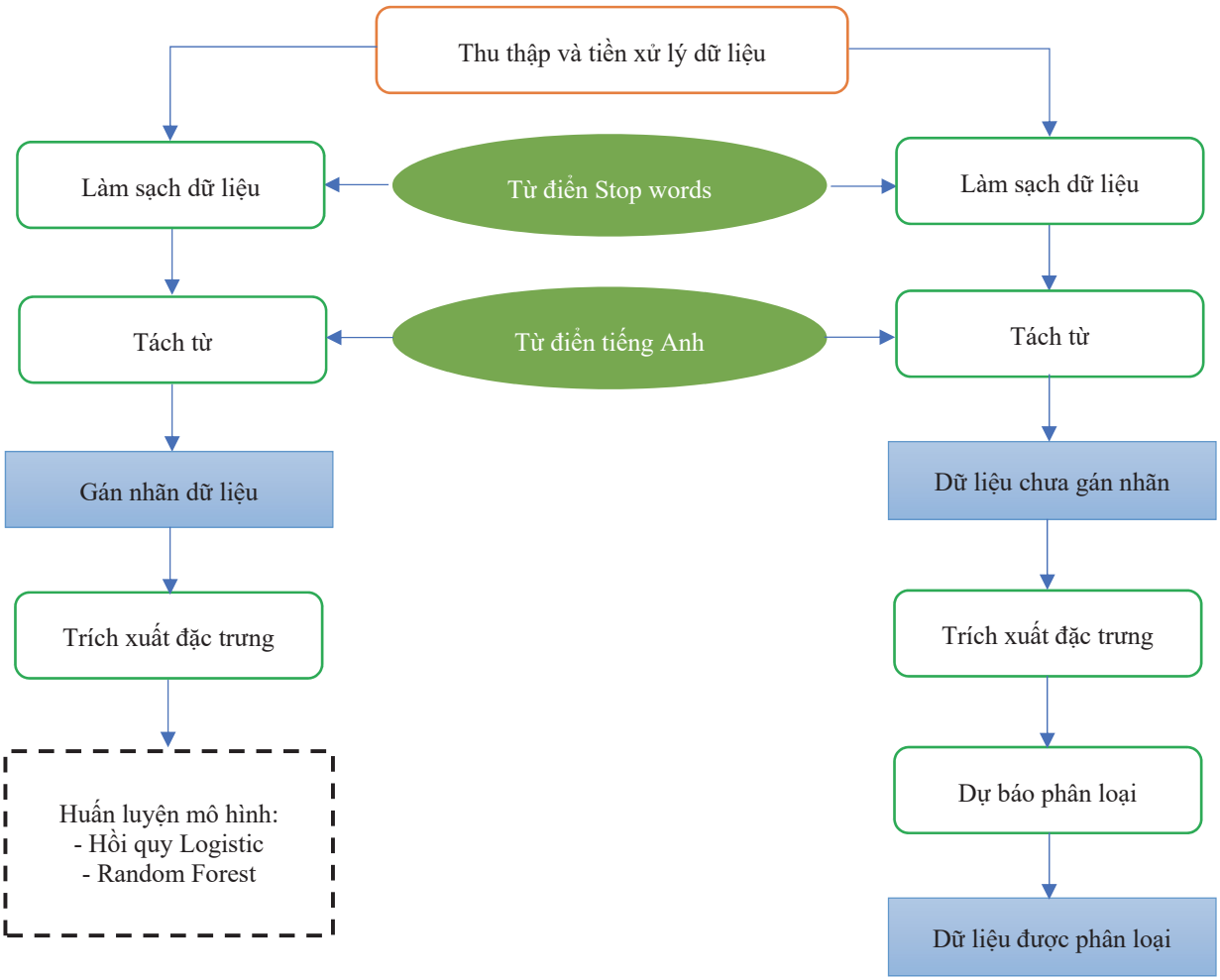
Phương pháp trên chỉ ra rằng một khu rừng bao gồm cây cối, càng có nhiều cây thì rừng càng lớn. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách lựa chọn tối ưu. Random forests có nhiều ứng dụng, chẳng hạn như ứng dụng trong xây dựng công cụ đề xuất, phân loại hình ảnh và lựa chọn tính năng. Bên cạnh đó, phương pháp có thể được sử dụng để phân loại các ứng viên cho vay trung thành, xác định hoạt động gian lận và dự đoán các bệnh (Nguyễn Duy Sim, 2018).

3. Phương pháp nghiên cứu

3.1. Mô hình nghiên cứu tổng quan

Để tiến hành thực nghiệm phương pháp phân loại các đánh giá, bình luận của khách hàng từ ứng dụng thương mại điện tử của Lazada trên Google Play dựa trên các phương pháp học máy có giám sát, Hình 3 dưới đây trình bày tổng quan các bước trong qui trình nghiên cứu.

Hình 3: Quy trình nghiên cứu



3.2. Thu thập dữ liệu

Trước tiên sử dụng các thư viện Selenium và Bs4 để thu thập tự động dữ liệu dưới dạng phi cấu trúc từ trang HTML Google Play có ứng dụng Lazada để truy cập vào môi trường API trên các ứng dụng và lưu thành các tập tin với định dạng JSON. Sau đó, chuỗi dữ liệu JSON được chuyển sang định dạng dữ liệu CSV và thực hiện phân tích rút trích trên tập dữ liệu thu được. Với đối tượng và phạm vi nghiên cứu hướng đến là phân tích ý kiến khách hàng được viết bằng ngôn ngữ tiếng Anh, do đó, dữ liệu chỉ sử dụng những ý kiến nhận xét, phản hồi của khách hàng về các sản phẩm, dịch vụ bằng ngôn ngữ tương ứng. Tổng số 53.321 ý kiến nhận xét của khách hàng đã được thu thập, một số thuộc tính được rút trích để phân tích bao gồm tên

người bình luận, điểm đánh giá, ngày đánh giá, nội dung bình luận, số lượng đồng tình với bình luận đó (lượt thích).

3.3. Tiền xử lý dữ liệu

Tiền xử lý là một điều quen thuộc đối với Khoa học dữ liệu. Đối với dữ liệu số, thông thường ta sẽ áp dụng một số quy tắc chuẩn hóa (nhằm giảm sự khác biệt giữa giá trị lớn nhất và nhỏ nhất), thay thế các giá trị không phải là dạng số (cũng như là các giá trị rỗng), phát hiện các giá trị ngoại lệ,... với sự hỗ trợ của bộ công cụ ngôn ngữ tự nhiên (Natural Language Toolkit - NLTK)[1] được công bố bởi Steven Bird, Edward Loper và cộng sự đến từ nhiều quốc gia. Bởi vì từ và cụm từ phức tạp hơn các số nguyên và số thực, dữ liệu cần được qua vài bước tiền xử lý hay còn gọi là luồn tiền xử lý (Preprocessing pipeline - PPL) bao gồm: Nhận diện ngôn ngữ, sử dụng thư viện Langdetect để xác định ngôn ngữ của đoạn text review, lựa chọn ngôn ngữ tiếng Anh và loại bỏ các ngôn ngữ khác nhằm giảm độ nhiễu của dữ liệu, tăng độ chính xác khi chạy models; Xử lý các biểu tượng emoji, sử dụng thư viện Emoji nhằm chuyển đổi các biểu tượng emoji thành text; Thay thế các từ viết sai, sử dụng công thức GOOGLETRANSLATE() của google sheets để nhận dạng các từ loại viết tắt, sai chính tả trong tiếng Anh, và sau đó thay thế chúng bằng từ có nghĩa; Tách từ (Tokenization), chuyển một dãy các ký tự thành một dãy các token (token là một dãy các ký tự mang ý nghĩa cụ thể, biểu thị cho một đơn vị ngữ nghĩa trong xử lý ngôn ngữ). Nhiều khi token được hiểu là một từ mặc dù cách hiểu này không hoàn toàn chính xác; Từ vựng hóa (Lemmatization), làm trở lại nguyên dạng ban đầu các từ vựng bị biến đổi thể (Inflection) hoặc được kết hợp (Conjugatetion). Ví dụ như biến đổi từ made thành make; Loại bỏ các từ dừng (Stop words), là loại bỏ những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Trong Tiếng Anh, stop words là những từ như: is, that, this... Mục đích của tác vụ này là loại bỏ những từ không mang lại nhiều ý nghĩa cho mô hình.

Sau giai đoạn tách từ, bước tiếp theo của phương pháp tiền xử lý mà bài báo thực hiện là Gán nhãn dữ liệu: mục đích này nhằm chuẩn bị tập dữ liệu đã được gán nhãn (hay đã được phân loại) đủ lớn để đưa vào làm tập dữ liệu huấn luyện. Dựa vào kết quả điểm đánh giá (ratings), nhận thấy các bình luận có điểm từ 1 đến 3 mang ý nghĩa tiêu cực (Negative), và ngược lại, các bình luận có điểm lớn hơn 3 mang ý nghĩa tích cực (Positive). Do đó, tập dữ liệu huấn luyện được xác định có 12.047 bình luận là tiêu cực (được gán nhãn 0) và 38.641 bình luận là tích cực (được gán nhãn 1); Trích xuất đặc trưng, chọn ra các đặc trưng tiêu biểu (chính là các từ khóa - Keywords) có tính đại diện cho tập dữ liệu để làm đầu vào cho thuật toán phân loại. Nghiên cứu này lựa chọn từ khóa theo phương pháp TF-IDF dựa trên thư viện Sklearn, giá trị TF-IDF của một từ khóa là thu được qua thống kê thể hiện mức độ quan trọng của từ khóa trong một bình luận. TF-IDF của từ khóa w_i trong bình luận d được tính bằng công thức sau:

$$tf\_idf_{id} = f_{id} \times log\frac{N}{n_i} \quad (8)$$

Trong đó: f_{id}: Tần suất xuất hiện của từ khóa w_i trong bình luận d

N: Tổng số bình luận

n_i : Số bình luận mà có từ khóa w_i xuất hiện.

Hình 4: Mẫu ma trận kết quả TF-IDF

m thuộc tính (từ)

n bình luận		using	year	helpful	handy	come	buying	...	great	place	shopping
	0	0,26	0,14	0,21	0,19	0,17	0,14	...	0	0	0
	...	...	...	...	...	...	...	...	...	...	...
	50687	0	0	0	0	0	0	...	0,77	0,49	0,40

Tại thư viện Sklearn.feature_extraction.text, tham số vector được khởi tạo bởi hàm TfidfVectorizer dưới dạng một biến có tên là vectorizer và sau đó gọi hàm fit_transform () để chuyển đổi danh sách các chuỗi thành ma trận biểu diễn các giá trị TF-IDF cho 50.688 (từ 0 đến 50.687) bình luận của khách hàng.

Ví dụ ở bình luận cuối cùng thể hiện 3 cột tương ứng với từ "great", "place", "shopping", có thể thấy từ "shopping" xuất hiện 10.673 lần trong toàn bộ tập dữ liệu và được biến đổi thành giá trị vector 0,4.

3.4. Huấn luyện mô hình

Bước tiếp theo là huấn luyện mô hình. Đây là giai đoạn quan trọng nhất của một mô hình khai phá ý kiến, nhằm mục đích xác định một bình luận của khách hàng là "tích cực" hay "tiêu cực". Nghiên cứu này ứng dụng hai thuật toán phân loại thuộc nhóm máy học giám sát là hồi quy Logistic và Random Forest, dựa trên kết quả tổng hợp từ các nghiên cứu trước có liên quan đến đề tài để tìm ra phương pháp phù hợp nhất đối với tập dữ liệu là các bình luận đã được phân loại. Từ đó, tiến hành dự báo cho các dữ liệu bình luận chưa được phân loại hoặc các dữ liệu bình luận mới phát sinh mà không cần phải huấn luyện lại.

Quá trình huấn luyện được tiến hành theo 2 cách:

- Phương pháp Hold-Out: toàn bộ tập dữ liệu sẽ được chia thành 2 tập con dữ liệu dùng để huấn luyện và dữ liệu dùng để kiểm thử không giao nhau: 70% tập huấn luyện, và 30% tập kiểm thử.

- Phương pháp K-Fold: hạn chế sự quá khớp (overfitting) để tối ưu giá trị và tăng độ chính xác cho mô hình, tập huấn luyện được chia thành K tập con không giao nhau có kích thước xấp xỉ nhau. Mỗi lần lặp một tập con trong K tập sẽ được dùng để làm tập kiểm thử, k-1 tập còn lại sẽ được sử dụng làm tập huấn luyện

Bảng 1: Ma trận nhầm lẫn

	Dự báo: Positive	**Dự báo: Negative**
Thực tế: Positive	True Positive (TP)	False Negative (FN)
Thực tế: Negative	False Positive (FP)	True Negative (TN)

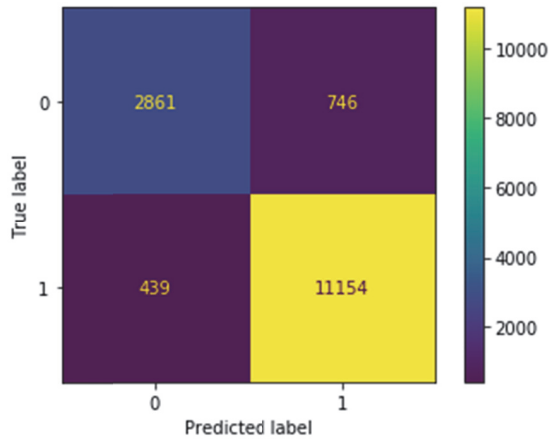
Đánh giá mô hình phân loại phổ biến dựa trên các chỉ số tính toán trong ma trận nhầm lẫn (Confusion Matrix) dựa trên nghiên cứu của Chicco & Jurman (2020).

Thông thường, hiệu quả của mô hình phân loại ý kiến được đánh giá dựa trên 4 chỉ số: Độ chính xác (Accuracy), Độ dự đoán (Precision), Độ bao phủ (Recall), và Giá trị trung bình điều hòa (F1). Ngoài ra, nghiên cứu này cũng xét đến yếu tố thời gian huấn luyện (Time) của từng mô hình.

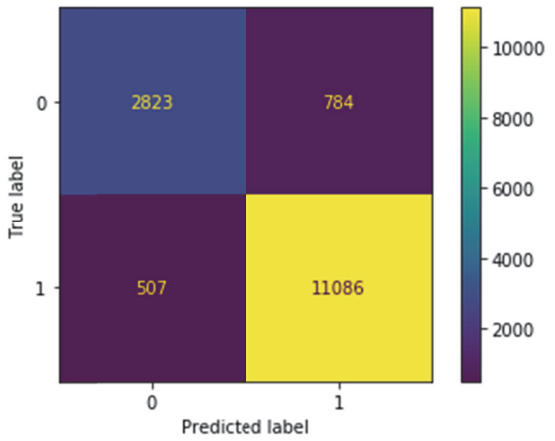
Trong đó:

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

Hình 5: Kết quả ma trận nhầm lẫn từ phương pháp hồi quy Logistic



Hình 6: Kết quả ma trận nhầm lẫn từ phương pháp Random Forest



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

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

$$F1 = \frac{2 \times Precision \times Recall}{Precision \times Recall} \quad (12)$$

Dựa vào kết quả ma trận nhầm lẫn của phương pháp hồi quy Logistic tại hình 5, có 2861 dữ liệu tích cực được dự báo đúng với tổng số lượng tích cực thực tế và 11154 dữ liệu tiêu cực được dự báo đúng với tổng lượng tiêu cực thực tế, cao hơn so với kết quả từ phương pháp Random Forest. Ngược lại, ma trận nhầm lẫn từ phương pháp Random Forest ở tại hình 6 có tới 784 dữ liệu tiêu cực và 507 dữ liệu tích cực là sai với dữ liệu thực tế, những dữ liệu này cần được loại bỏ để tránh gây nhiễu cho mô hình.

4. Kết quả và thảo luận

4.1. Đánh giá mô hình

Bảng 2: Kết quả mô hình theo phương pháp Hold-Out

Tập dữ liệu	Các độ đo	Bình luận	Hồi quy Logistic	Random Forests
Huấn luyện	Accuracy		0,94	1
	Precision	0	0,90	1
		1	0,95	1
	Recall	0	0,83	0,99
		1	0,97	1
	F1-score	0	0,86	0,99
		1	0,96	1
Kiểm thử	Accuracy		0,92	0,92
	Precision	0	0,87	0,85
		1	0,94	0,93
	Recall	0	0,79	0,78
		1	0,96	0,96
	F1-score	0	0,83	0,81
		1	0,95	0,94

Sau khi tiến hành nghiên cứu thực nghiệm mô hình theo hai phương pháp Hold-Out và K-Fold bằng các thuật toán hồi quy Logistic, Random Forest dựa trên tập dữ liệu huấn luyện và kiểm thử với kết quả tại Bảng 2.

Đối với tập huấn luyện, theo phương pháp Hold-Out thì độ chính xác (Accuracy) của thuật toán hồi quy Logistic thấp hơn Random Forest khoảng 5%, như vậy độ chính xác của Random Forest gần như tuyệt đối 100%, nhưng về tập dữ liệu kiểm thử thì cả hai phương pháp có cùng độ chính xác là 92%.

Về hệ số dự đoán (Precision), thuật toán hồi quy Logistic dự đoán chính xác khoảng 95% đối với dự đoán tích cực ở cả hai tập dữ liệu, có nghĩa là trong 100 dữ liệu tích cực thực tế thì mô hình dự đoán đúng 95 dữ

Bảng 3: Kết quả mô hình theo phương pháp K-Fold (K=5)

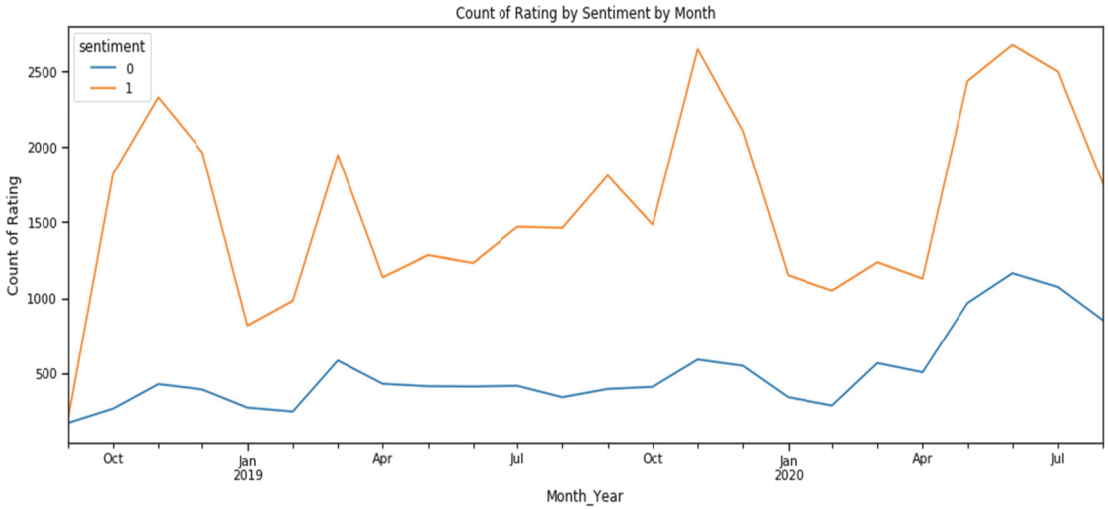
K	Hồi quy Logistic	Random Forests
1	0,918	0,915
2	0,924	0,918
3	0,919	0,916
4	0,923	0,913
5	0,921	0,916
Average	***0,921***	***0,916***

liệu tích cực. Tương tự, Random Forest lại cho ra độ dự đoán chính xác, độ bao phủ (Recall), giá trị trung bình điều hòa giữa Recall và Precision (F1-score) ở tập huấn luyện có giá trị cao hơn hồi quy Logistic, tuy nhiên xét về khía cạnh thử nghiệm, hồi quy Logistic là sự lựa chọn chính xác và hiệu quả hơn.

Kết quả mà phương pháp K-Fold mang lại cho thấy độ chính xác trung bình vòng lặp K ở hai mô hình tương đương bằng nhau khoảng 92%. Nghĩa là hai mô hình khác nhau theo hai phương pháp khác nhau chứng tỏ được rằng khi đưa dữ liệu kiểm thử vào cho ra độ chính xác gần khớp nhau và tương đối phù hợp trong các ứng dụng tiếp theo để phân loại ý kiến cho các dữ liệu bình luận chưa được phân loại hoặc các dữ liệu bình luận mới phát sinh.

4.2. Trực quan hóa kết quả thực nghiệm

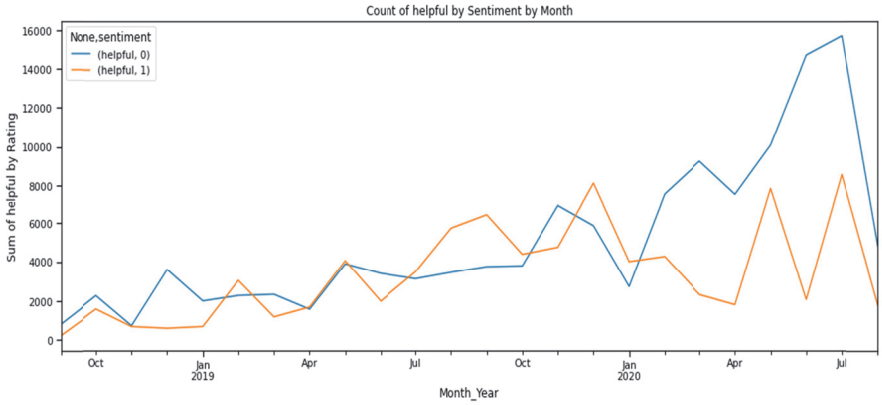
Hình 7: Biểu đồ đường thể hiện số lượng đánh giá theo tháng



Sau khi thực thi mô hình, nhằm dễ dàng theo dõi và ra quyết định, kết quả thực nghiệm được trực quan hóa. Hình 7 biểu thị kết quả phân tích số bình luận dựa theo từng giai đoạn thời gian (tháng).

Đồ thị tại Hình 7 diễn tả tỉ lệ người có bình luận tiêu cực về ứng dụng Lazada cao nhất vào tháng 3 của năm 2019 và tháng 6 năm 2020, tuy nhiên, nhìn một cách tổng quan thì những bình luận tiêu cực đang có xu

Hình 8: Biểu đồ đường thể hiện số lượng đồng tình với các đánh giá theo tháng

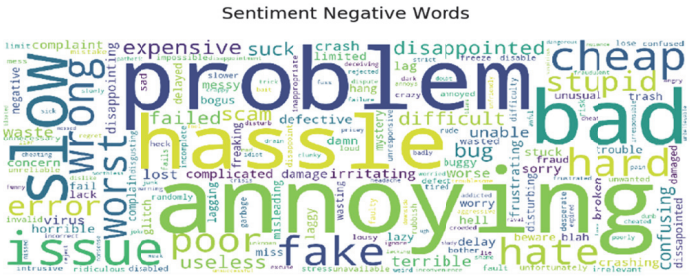


hướng tăng. Và ngược lại, đối với bình luận tích cực luôn chiếm vị trí cao hơn các bình luận tiêu cực, dao động đều qua các tháng và đạt tới điểm cao nhất vào tháng 6 năm 2020 khoảng trên 2.800 bình luận, cho thấy số người vào truy cập và đánh giá trên trang trong mùa dịch covid vừa qua gia tăng đáng kể.

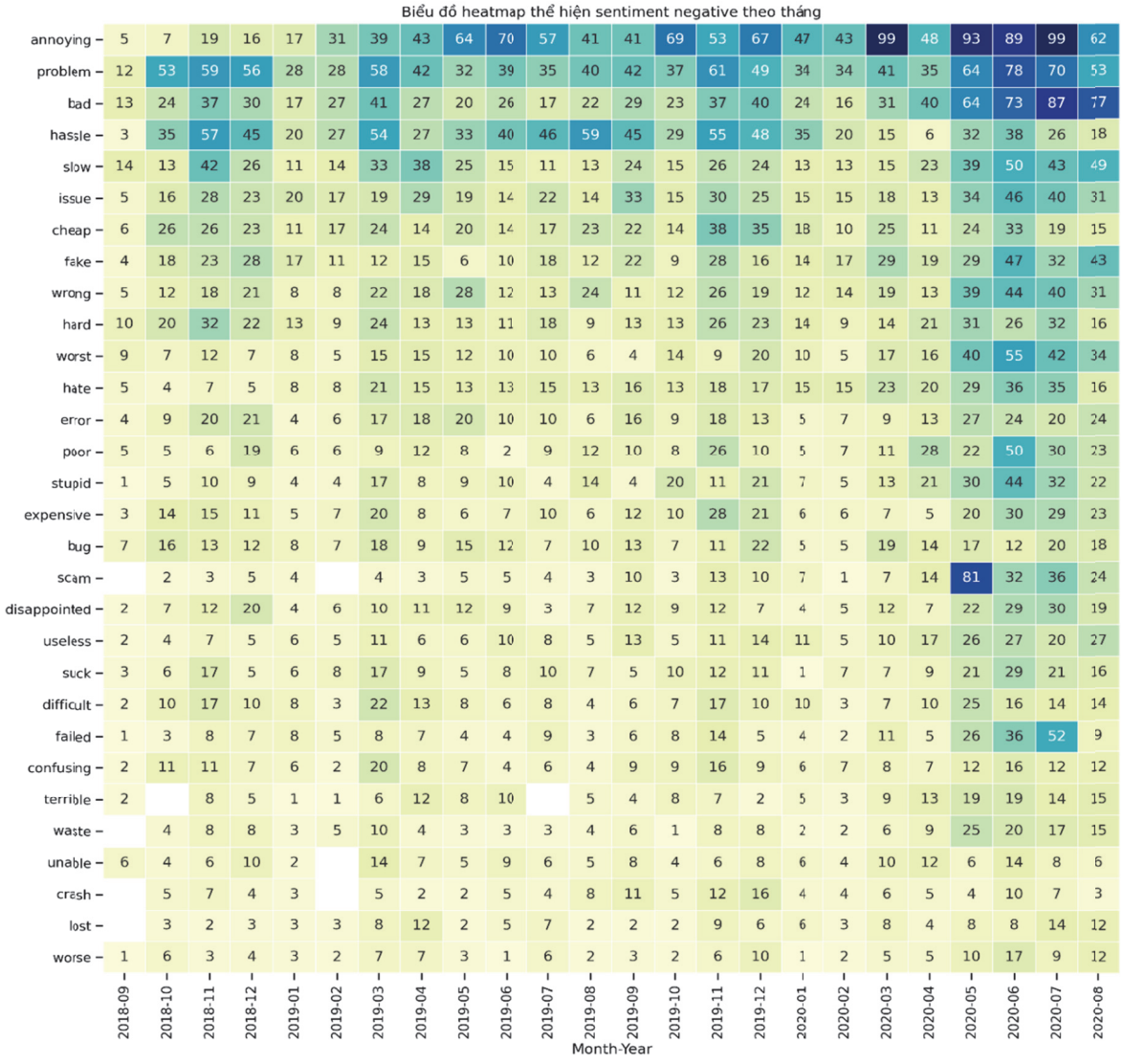
Bên cạnh đó, tất cả mọi người sử dụng dịch vụ Lazada mang lại dường như luôn đưa ra quan điểm đồng ý

với ý kiến tiêu cực của người đánh giá. Biểu đồ đường ở Hình 8 cho thấy cảm xúc của người tiêu dùng đánh giá một cách chân thật, đúng đắn và khách quan về mặt tiêu cực sẽ nhận được nhiều sự tán thành hơn là các

Hình 9: Keywords thể hiện cảm xúc tiêu cực



Hình 10: Biểu đồ heatmap thể hiện số lượng các từ tiêu cực theo tháng

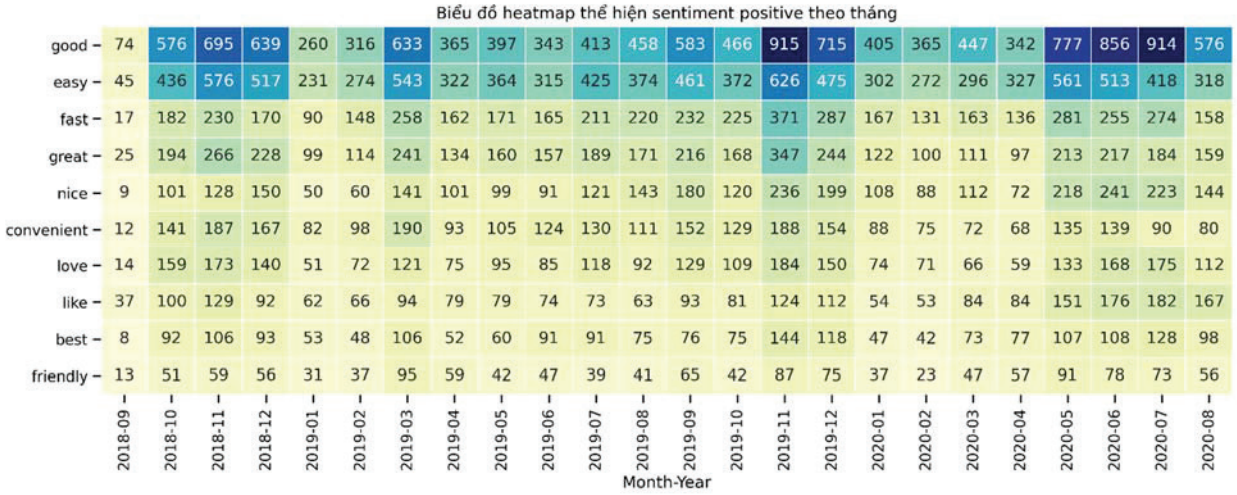


đánh giá tích cực vì sẽ có những người có thể xem việc đánh giá như một việc không quan trọng, phiền phức đối với họ. Cụ thể có tới gần 16.000 người đồng quan điểm với đánh giá tiêu cực vào tháng 7 năm 2020 và đang có xu hướng tăng.

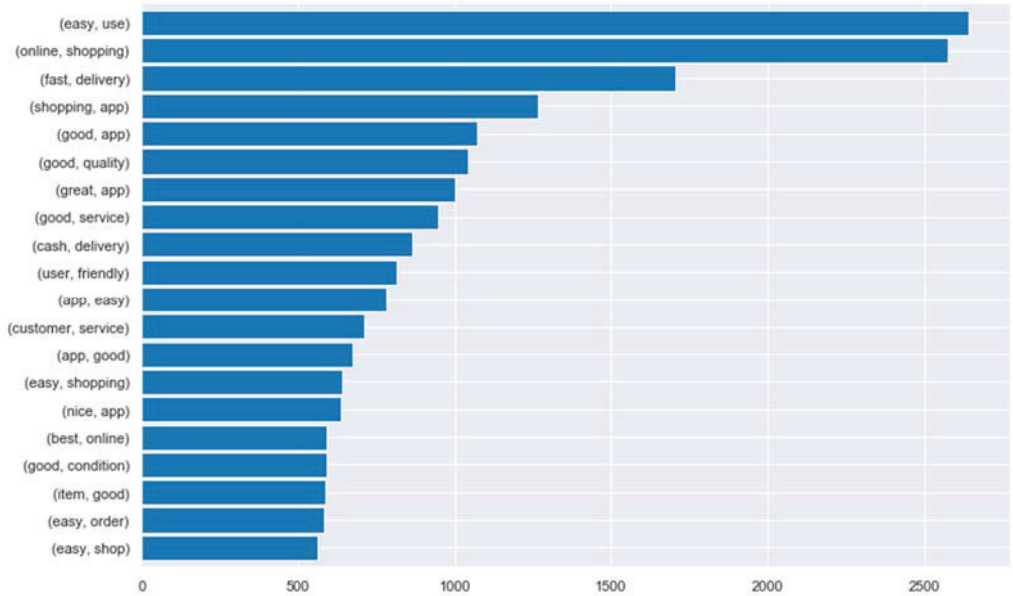
Trong những tháng giữa năm 2020, số từ bình luận tiêu cực tăng đột biến so với những tháng cuối năm 2018. Theo Hình 10, nổi bật ở những từ "annoying", "bad", "scam" ngày càng xuất hiện nhiều trong các

Hình 11: Keywords thể hiện cảm xúc tích cực

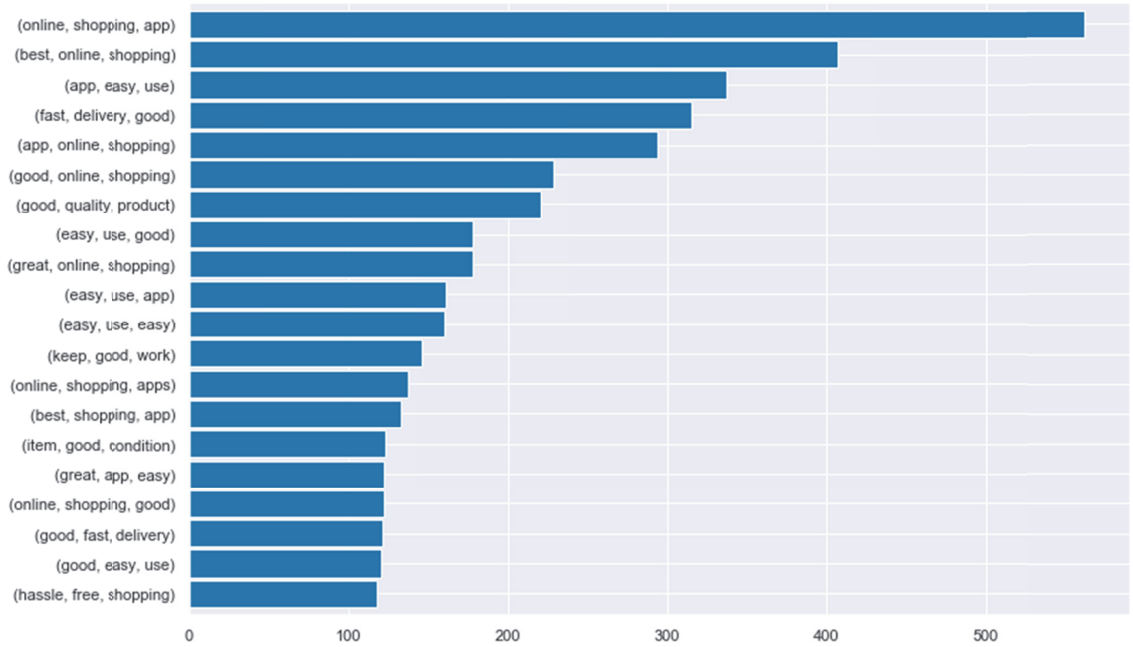
Hình 12: Biểu đồ heatmap thể hiện số lượng các từ tích cực theo tháng



Hình 13: Đồ thị thanh ngang thể hiện ngẫu nhiên 20 bình luận có chứa 2 từ



Hình 14: Đồ thị thanh ngang thể hiện ngẫu nhiên 20 bình luận có chứa 3 từ



đánh giá nhận xét của khách hàng chứng tỏ sau một thời gian trải nghiệm và sử dụng mọi sản phẩm, dịch vụ thông qua Lazada, khách hàng mới dần cảm nhận rồi từ đó nêu ra quan điểm thực tế hơn. Các nhà điều hành cần phải xem xét, thống kê lại một số những bình luận chứa các từ trên để biết chính xác khách hàng đang thảo luận về sản phẩm, dịch vụ nào từ tháng 3 đến tháng 7 năm 2020 để cải thiện lại tình hình hoạt động kinh doanh ngày càng hiệu quả hơn, cũng là một trong những cách để hiểu rõ hơn về hành vi người tiêu dùng, nhằm giữ chân khách hàng để giảm tối đa tỉ lệ khách hàng rời bỏ để đến những sàn thương mại của các đối thủ khác.

Nhìn chung những keywords được thể hiện thông qua Hình 11 và Hình 12 mang ý nghĩa tích cực gấp khoảng 10 lần các từ tiêu cực và dao động khá mạnh qua các tháng, chi tiết ở những từ như "good" – gần 1.000 từ, "easy" – khoảng 500 từ chiếm phần lớn trong số các đánh giá. Đa phần, hiệu suất kinh doanh của các sàn thương mại điện tử hàng đầu luôn mang lại hiệu quả khá cao đến với người tiêu dùng sau vài năm dẫn đầu thị trường về lượng người truy cập vì họ luôn biết cải tiến, luôn sẵn sàng phát huy, sáng tạo không ngừng để đáp ứng nhu cầu khách hàng theo thời gian. Một lý do khác là có thể những từ này mang ý nghĩa ngắn gọn, xúc tích, tiện lợi để người dùng đưa ra đánh giá nhanh chóng bao gồm cả những người không quan trọng việc đánh giá.

Hình 13 và 14 dùng thuật toán ngrams trong thư viện NLTK[2] trên tập 20 dữ liệu bình luận ngẫu nhiên để thấy rõ được những từ mang ý nghĩa tốt đẹp là thực sự theo hướng tích cực, hầu như các từ này không thuộc trong câu phủ định hay nghi vấn của khách hàng vì không có từ "not", "yet", đa phần các bình luận của khách hàng là câu khẳng định.

5. Kết luận

Dựa vào những kết quả đã nghiên cứu và phân tích từ tập dữ liệu lớn về ứng dụng Lazada trên trang web Google Play trong khoảng thời gian từ năm 2018 đến năm 2020, bài báo đã nghiên cứu và lựa chọn được mô hình phân loại ý kiến phù hợp với dữ liệu trên và có kết quả thử nghiệm với độ chính xác cao và có ý nghĩa thực tiễn. Đây được xem là bước quan trọng nhất của quy trình khai phá ý kiến, làm nền tảng cho việc ứng dụng khai phá ý kiến trong nhiều lĩnh vực. Hơn thế nữa, quá trình phân tích, kết quả, mô hình đề xuất có thể nhận biết được cảm xúc tích cực và tiêu cực trong ý kiến khách hàng đã ảnh hưởng 90% đến ứng dụng Lazada, điển hình tập trung vào các tính năng, chất lượng của ứng dụng. Từ đó, dựa vào kết quả nghiên cứu và trực quan hóa, những thông tin và tri thức có được sẽ giúp nhà quản trị khắc phục được những vấn đề mà khách hàng thường quan tâm một cách hiệu quả và tốt hơn tuy không phải toàn bộ nhưng phần nào cũng nắm rõ những vấn đề nóng liên quan đến trải nghiệm của khách hàng trên ứng dụng.

Tuy nhiên, mô hình nghiên cứu còn một số hạn chế, chưa phân tích dữ liệu trong thời gian thực, thời gian huấn luyện và thời gian dự đoán để nhà quản trị thương mại điện tử đưa ra quyết định nhanh chóng. Ngoài ra, nghiên cứu sẽ tiếp tục hướng đến các mục tiêu: thứ nhất, không chỉ dừng lại ở tập dữ liệu này mà còn mở rộng thêm việc khai phá, trích xuất dữ liệu lớn từ các trang mạng xã hội cũng như trang thương mại điện tử khác một cách tự động hóa và ứng dụng hai mô hình phân loại đã nghiên cứu cho ra độ chính xác tương đối cao vào giải quyết. Thứ hai, song song việc thu thập và xử lý dữ liệu, nghiên cứu đề xuất phương pháp tạo ra những báo cáo trực quan về phân tích cảm xúc người dùng vào đúng thời gian thực tế mà các nhà chiến lược kinh doanh mong muốn, giúp phục vụ nhu cầu cho việc đưa ra quyết định kịp thời để nâng cao sự hài lòng của khách hàng. Cuối cùng là ngoài việc phân tích ý kiến, đánh giá thành hai yếu tố tích cực, tiêu cực, bài báo cũng tiếp tục nghiên cứu giải pháp phân loại cảm xúc thành nhiều mức độ khác nhau, trong đó có cả bình luận mang yếu tố trung lập để dự đoán chính xác hơn hành vi, sở thích người tiêu dùng.

Ghi chú:

1, 2. NLTK (2020), *Natural Language Toolkit*, retrieved on October 22nd 2020, from <https://www.nltk.org/>.

Tài liệu tham khảo

Ahuja, R., Chug, A., Kohli, K., Gupta, S. & Ahuja, P. (2019), 'The impact of features extraction on the sentiment analysis', *Procedia Computer Science*, 152(2019), 341-348.

Bahrawi, B. (2019), 'Sentiment analysis using random forest algorithm online social media based', *Journal of Information Technology and Its Utilization*, 2(2), 29-33.

Chicco, D. & Jurman, G. (2020), 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 6(21), retrieved on October 22nd 2020, from <https://doi.org/10.1186/s12864-019-6413-7>.

Đỗ Minh Hải (2017), *Hồi quy logistic (Logistic Regression)*, truy cập ngày 22 tháng 12 năm 2020, từ <https://dominhhai.github.io/vi/2017/12/ml-logistic-regression/>.

Ho, T.K. (1995), 'Random decision forest', *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, 278-282.

Hussein, Doaa Mohey El-Din Mohamed (2016), 'Analyzing scientific papers based on sentiment analyses, Thesis of master, Faculty of computers and information, Cairo University.

Kaushik, A., Kaushik, A. & Naithani, S. (2015), 'A study on sentiment analysis: Methods and tools', *International Journal of Science and Research*, 4(12), 287-292.

Le, H.S., Trieu, C., Ho, T., Lee, J.H. & Lee, H.K. (2017), 'Applying artificial neural network for sentiment analytics of social media text data in fastfood industry', *The Journal of Internet Electronic Commerce Resarch*, 17(5), 113-123.

Mali, D., Abhyankar, M., Bhavarthi, P., Gaidhar, K. & Bangare, M. (2016), 'Sentiment analysis of product reviews for E-commerce recommendation', *International Journal of Management and Applied Science*, 2(1), 127-131.

Nguyễn Duy Sim (2018), *Phân lớp bằng Random Forests trong Python*, truy cập ngày 22 tháng 12 năm 2020, từ <https://viblo.asia/p/phan-lop-bang-random-forests-trong-python-djeZ1D2QKWz>.

Pham, H.O. (2018), 'Introduction to NLP and deep learning', *Natural Language Processing with Deep Learning*, retrieved on October 22nd 2020, from <https://viblo.asia/p/lecture-1-introduction-to-nlp-and-deep-learning-bJzKmXzk59N>.

Salmiah, S., Sudrajat, D., Nasrul, N., Agustin, T., Harani, N.H. & Nguyen, P.T. (2019), 'Sentiment Analysis for Amazon Products using Isolation Forest', *International Journal of Engineering and Advanced Technology (IJEAT)*, 8(6S), 894-897.

Schuster, M. & Paliwal, K.K. (1997), 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing*, 45(11), 2673 - 2681.

Trần Thị Ngọc Thảo, Nguyễn Ngọc Kim Liên & Ngô Minh Vương (2014), 'Phân tích ý kiến của nhận xét tiếng anh dựa trên phương pháp học máy', *Tạp chí Khoa học và Công nghệ*, 52(4D), 142-155.