

Mô hình khám phá trải nghiệm khách hàng dựa trên phương pháp phân tích quan điểm và máy học

Nguyễn Trần Thúy Quỳnh^{1,2}, Bùi Nguyễn Bích Ngọc^{1,2}, Nguyễn Thị Bảo Trâm^{1,2}, Trần Nhật Nguyên^{1,2},
Võ Bá Tùng^{1,2}, Hồ Trung Thành^{1,2,*}



Use your smartphone to scan this
QR code and download this article

TÓM TẮT

Thấu hiểu tâm lý khách hàng là một trong những vấn đề trọng tâm trong chiến lược phát triển lâu dài của doanh nghiệp, giúp xây dựng niềm tin với khách hàng và nâng cao lợi thế cạnh tranh của doanh nghiệp trong thị trường. Đặc biệt, với sự phát triển mạnh mẽ của công nghệ, số lượng người dùng trực tuyến tăng nhanh, những tương tác và bình luận trên mạng trở thành tập dữ liệu thực tế vô cùng quan trọng, hỗ trợ khai thác hành vi và tâm lý của người dùng. Bài nghiên cứu hướng đến mục tiêu phân tích quan điểm tích cực hay tiêu cực của khách hàng thông qua các bình luận trong lĩnh vực thương mại điện tử dựa trên sự kết hợp giữa phương pháp máy học có giám sát và kỹ thuật xử lý ngôn ngữ tự nhiên; đồng thời, đánh giá hiệu suất mô hình qua ma trận nhầm lẫn nhằm đảm bảo độ tin cậy của kết quả. Bài nghiên cứu thực nghiệm mô hình đề xuất với 04 phương pháp máy học bao gồm Naive Bayes, SVM, Random Forest và Hồi quy Logistic trên 935.000 bình luận được thu thập từ 04 ứng dụng thương mại điện tử phổ biến tại Việt Nam (Tiki, Shopee, Lazada và Sendo). Kết quả nghiên cứu đã phân loại quan điểm tích cực và tiêu cực với độ chính xác cao trong khoảng 91%, thể hiện qua những báo cáo, biểu đồ phản ánh xu hướng và tâm lý của khách hàng. Bài nghiên cứu cũng mang đến những góc nhìn mới, sâu sắc hơn về hành vi khách hàng, giúp nhà quản trị tìm ra điểm mạnh, điểm yếu của dịch vụ cũng như ứng dụng, từ đó cải thiện và nâng cao trải nghiệm người dùng. Các doanh nghiệp thương mại điện tử có thể ứng dụng trong phân tích xu hướng thị trường, nắm bắt nhu cầu, mong muốn của khách hàng để xây dựng chiến lược phát triển sản phẩm và dịch vụ một cách hiệu quả.

Từ khóa: Phân tích quan điểm, máy học, mô hình phân loại, bình luận, thương mại điện tử

¹Trường Đại học Kinh tế - Luật

²Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

Liên hệ

Hồ Trung Thành, Trường Đại học Kinh tế - Luật

Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

Email: thanhht@uel.edu.vn

Lịch sử

- Ngày nhận: 20-3-2022
- Ngày chấp nhận: 09-9-2022
- Ngày đăng:

DOI :



Bản quyền

© ĐHQG Tp.HCM. Đây là bài báo công bố mở được phát hành theo các điều khoản của the Creative Commons Attribution 4.0 International license.



1 GIỚI THIỆU

Với sự phát triển mạnh mẽ của cuộc cách mạng công nghệ di động được dẫn đầu bởi các điện thoại thông minh và các phần mềm ứng dụng, quá trình phát triển hoạt động bán lẻ đã được thúc đẩy tạo nên sự dịch chuyển tương tác của người tiêu dùng và nhà bán lẻ. Theo báo cáo “Kinh tế Internet khu vực Đông Nam Á năm 2020” của Google, Temasek và Bain & Company, số lượng người truy cập Internet từ năm 2015-2020 tăng từ 260 đến 400 triệu trong năm 2020, trong đó tỷ lệ người mua sắm trực tuyến mới tại Việt Nam đạt 41% cho thấy sự thay đổi hành vi khách hàng, xu hướng mua sắm trực tuyến tăng lên và là dấu hiệu lạc quan cho các doanh nghiệp đầu tư phát triển thương mại điện tử. Trong năm 2020, Thủ tướng Chính Phủ phê duyệt quyết định 645/QĐ-TTg về Kế hoạch tổng thể phát triển thương mại điện tử quốc gia giai đoạn 2021-2025. Quyết định này cho thấy doanh nghiệp là lực lượng nòng cốt triển khai thương mại điện tử với mục tiêu mở rộng thị trường tiêu thụ và thúc đẩy ứng dụng rộng rãi thương mại điện tử cho các doanh nghiệp.

Khi người tiêu dùng sử dụng các thiết bị di động ưu tiên mua sắm trực tuyến để tìm kiếm thông tin, truy cập, so sánh, đánh giá các sản phẩm thì giải pháp thương mại điện tử là kênh kinh doanh đóng vai trò quan trọng đối với doanh nghiệp. Các doanh nghiệp, nhà bán hàng tăng cường các chiến lược kỹ thuật số nhằm tiếp cận người dùng hiệu quả và đây là cơ hội để thu hút khách hàng, hiện diện thương hiệu doanh nghiệp trên các nền tảng trực tuyến. Bên cạnh đó, thấu hiểu tâm lý khách hàng để đưa chiến lược tiếp thị phù hợp, xây dựng lòng tin đối với họ càng trở nên quan trọng hơn bao giờ hết. Việc thu thập, phân tích các đánh giá của người dùng trên các nền tảng di động khi trải nghiệm mua sắm trên các nền tảng kỹ thuật số với mục đích nâng cao trải nghiệm khách hàng để từ đó gia tăng sự trung thành, hài lòng và nâng cao được chất lượng và phát triển ứng dụng tốt hơn trong môi trường cạnh tranh như hiện nay^{1,2}. Chính vì vậy, phân tích quan điểm của khách hàng dựa trên những bình luận và phản hồi trên các kênh mua sắm trực tuyến là vô cùng quan trọng. Việc khai thác dữ liệu một cách hiệu quả giúp doanh nghiệp biết được trải

Trích dẫn bài báo này: Quỳnh N T T, Ngọc B N B, Trâm N T B, Nguyên T N, Tùng V B, Thành H T. **Mô hình khám phá trải nghiệm khách hàng dựa trên phương pháp phân tích quan điểm và máy học.** *Sci. Tech. Dev. J. - Eco. Law Manag.*; (1):1-13.

45 nghiệm của khách hàng, nắm bắt những vấn đề đang
46 gặp phải nhằm cải thiện kết quả kinh doanh và đưa
47 ra những chiến lược thị trường tốt hơn, giúp cắt giảm
48 chi phí, tăng doanh thu.

49 Nội dung tiếp theo của bài báo là phần 2 gồm cơ sở lý
50 thuyết và các nghiên cứu liên quan, nhằm định hình,
51 xác định các mô hình, phương pháp nghiên cứu phù
52 hợp với mục tiêu đặt ra. Các vấn đề liên quan và quá
53 trình thực nghiệm được mô tả trong phần 3 - phương
54 pháp và quy trình thực hiện nghiên cứu. Sau quá trình
55 thực nghiệm, kết quả thực nghiệm được tìm ra và
56 những thảo luận được đề cập trong phần 4. Phần cuối
57 cùng là kết luận và hướng phát triển của nghiên cứu.

58 CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN 59 CỨU LIÊN QUAN

60 Phân tích quan điểm

61 Phân tích quan điểm còn được gọi là khai thác ý kiến,
62 là lĩnh vực nghiên cứu về ý kiến, tình cảm, đánh giá,
63 thẩm định, thái độ và quan điểm đối với các thực thể
64 như sản phẩm và dịch vụ^{3,4}. Phân tích quan điểm
65 thường được phân loại thành 03 khía cạnh: tích cực,
66 tiêu cực và trung tính. Phân tích quan điểm được áp
67 dụng để quản lý thông tin chính phủ⁵ cho phép chính
68 phủ theo dõi được những ý kiến đóng góp hoặc phản
69 ánh của người dân; ứng dụng vào phân tích những
70 tin tức thời sự nhằm phân tích những nội dung tin
71 tức hoặc xác định xu hướng tin tức được quan tâm
72 nhiều nhất⁶; ứng dụng trong lĩnh vực giao đồ ăn trực
73 tuyến để phân tích những phản hồi của khách hàng
74 về sự hài lòng của họ nhằm cải thiện và nâng cao chất
75 lượng dịch vụ và trải nghiệm người dùng⁷. Ngoài ra,
76 phân tích quan điểm cũng được áp dụng để cải thiện
77 hệ thống giáo dục dựa vào các đánh giá về khóa học,
78 học kỳ hay thậm chí là giảng viên giảng dạy⁸.

79 Đối với một doanh nghiệp, những nội dung do người
80 dùng tạo trên các ứng dụng di động đã có thể cung
81 cấp cho doanh nghiệp những thông tin về ý kiến tích
82 cực, tiêu cực hay trung tính của người tiêu dùng về sản
83 phẩm của doanh nghiệp và của đối thủ cạnh tranh⁹.
84 Các doanh nghiệp thường gặp khó khăn trong việc
85 đo lường mức độ quan tâm của người tiêu dùng và
86 xác định dữ liệu người dùng nào thực sự hữu ích để
87 họ thu thập. Bằng cách sử dụng phân tích quan điểm
88 được bổ sung với sự thông minh của con người, các
89 doanh nghiệp có thể lọc ra dữ liệu nhiễu, mơ hồ và với
90 sự trợ giúp của công nghệ máy học có thể xác định dữ
91 liệu quan trọng thúc đẩy hoạt động kinh doanh của
92 họ¹⁰.

Các phương pháp phân tích quan điểm

Cho đến hiện nay, có nhiều phương pháp được nghiên
cứu và áp dụng vào phân tích ý kiến khách hàng trực
tuyến. Trong đó, điển hình là hai phương pháp đó
là phương pháp máy học (machine learning)^{11,12} và
phương pháp dựa trên từ vựng (lexicon-based)¹³⁻¹⁵.
Mỗi phương pháp có những ưu điểm và hạn chế khi
được áp dụng. Chẳng hạn như, phương pháp dựa trên
từ vựng là một cách tiếp cận không giám sát tương đối
dễ thực hiện¹⁶. Tuy nhiên, đối với các tập dữ liệu lớn
với nhiều đặc trưng để phân tích thì chỉ phương pháp
dựa trên từ vựng là không đủ để tiếp cận một cách
có hiệu quả^{17,18}. Trong khi đó các phương pháp máy
học có giám sát khi phân loại dữ liệu thành các lớp yêu
cầu dữ liệu đầu vào phải sạch và được gán nhãn theo
cấu trúc nhất định¹⁹. Dữ liệu thô phải được xử lý,
lấy mẫu, dán nhãn; với ba tập dữ liệu mẫu gồm huấn
luyện, xác thực và kiểm tra⁷.

Các phương pháp máy học được áp dụng cho phân
tích quan điểm chủ yếu thuộc về phân loại có giám
sát. Một số phương pháp máy học được sử dụng để
phân loại các đánh giá: Naive Bayes²⁰, Hồi quy Logistic²¹,
Random Forest²², Support Vector Machine²³
và nhiều mô hình khác. Phương pháp máy học bắt
đầu từ việc thu thập tập dữ liệu huấn luyện, trích xuất
đặc trưng²⁴ và sau đó huấn luyện một bộ phân loại
trên dữ liệu huấn luyện. Cuối cùng, phân loại có giám
sát cho biết cách thức tập dữ liệu thể hiện²⁵.

Các nghiên cứu trước đó đã đạt được những kết quả
tốt trong lĩnh vực phân tích quan điểm, tuy nhiên
vẫn còn những hạn chế như trao đổi trên. Trong
nghiên cứu này, chúng tôi đề xuất kết hợp phương
pháp máy học và kỹ thuật xử lý ngôn ngữ tự nhiên
(hybrid method)¹¹ nhằm nâng cao tính chính xác và
hiệu quả trong phân tích quan điểm và đồng thời thực
nghiệm mô hình trên bộ dữ liệu thực tế (những bình
luận, ý kiến phản hồi của khách hàng trên 04 ứng
dụng thương mại di động). Từ kết quả thực nghiệm,
nghiên cứu sẽ áp dụng phương pháp ma trận nhầm
lẫn để đánh giá hiệu suất của phương pháp đề xuất sau
khi huấn luyện và lựa chọn mô hình hiệu quả nhất để
áp dụng cho dữ liệu thực tế và đưa ra phương pháp
phù hợp giúp cho các nhà phát triển ứng dụng có thể
xác định những ưu điểm hiện có của ứng dụng để tiếp
tục phát huy, đồng thời nhận ra những khuyết điểm
còn tồn tại để nhanh chóng khắc phục, nâng cao tối
đa trải nghiệm mua sắm trực tuyến của người dùng.

PHƯƠNG PHÁP VÀ MÔ HÌNH NGHIÊN CỨU ĐỀ XUẤT

Phương pháp nghiên cứu

Trong nghiên cứu này, sự kết hợp hai phương pháp
khám phá ý kiến khách hàng dựa trên từ vựng và

phương pháp máy học. Trong đó, phương pháp dựa trên từ vựng phù hợp để phân tích câu và mức độ đặc trưng mà không cần dữ liệu huấn luyện. Phương pháp máy học sử dụng các yếu tố ngôn ngữ để giải quyết phân loại quan điểm và đưa ra dự đoán. Khai thác ưu điểm của cả hai phương pháp này bằng phương pháp lai kết hợp (hybrid method), chúng tôi đã đề xuất một mô hình phân tích quan điểm bằng phương pháp lai và thực nghiệm mô hình trên bộ dữ liệu trong lĩnh vực thương mại di động tại Việt Nam.

Mô hình nghiên cứu đề xuất

Hình 1 trình bày mô hình nghiên cứu khám phá trải nghiệm khách hàng trong lĩnh vực thương mại di động dựa trên phương pháp phân tích quan điểm và học máy. Trong đó, quy trình thực hiện nghiên cứu được tích hợp vào mô hình để tiến hành từ bước phân tích yêu cầu, thu thập dữ liệu thô các đánh giá mà khách hàng để lại từ 04 ứng dụng thương mại di động Tiki, Shopee, Sendo và Lazada. Tập dữ liệu này được tiền xử lý, chuẩn hóa và gán nhãn trước khi đưa vào trích xuất đặc trưng. Bộ dữ liệu của 04 ứng dụng thương mại di động được chia làm hai phần: tập dữ liệu huấn luyện (training data) và tập dữ liệu kiểm tra (test data). Tập dữ liệu huấn luyện sử dụng để thiết lập các phương pháp máy học và tập dữ liệu kiểm tra được dùng để đánh giá các phương pháp máy học, từ đó chọn ra mô hình phù hợp nhất với bộ dữ liệu thu thập được. Cuối cùng, sau khi có được mô hình phù hợp với bộ dữ liệu, dữ liệu được trực quan hóa để có thể so sánh, đánh giá giữa các ứng dụng thương mại di động và tìm ra hướng đi phù hợp cho các doanh nghiệp triển khai thương mại di động.

Thu thập dữ liệu

Để thực nghiệm phương pháp đề xuất, chúng tôi đã thực hiện việc nghiên cứu để tải, tiến hành thu thập hơn 1.000.000 bình luận từ 04 ứng dụng thương mại di động lớn tại Việt Nam bao gồm Shopee, Lazada, Tiki và Sendo. Bảng 1 trình bày các trường dữ liệu được thu thập từ 04 ứng dụng thương mại di động. Theo thứ tự từ trái qua phải, trường “userName” là cột dữ liệu về tên khách hàng/người bình luận, “content” là cột dữ liệu trình bày nội dung bình luận, “at” là cột dữ liệu về thời gian khách hàng đó bình luận và được đăng tải lên, “score” là cột dữ liệu thể hiện điểm đánh giá từ 1 đến 5 cho ứng dụng, “thumbsUpCount” là cột dữ liệu thể hiện số lượt thích của những người khác dành cho bình luận đó, “address” là cột dữ liệu về tên ứng dụng mà người đó bình luận.

Tiền xử lý dữ liệu

Trước tiên, dữ liệu thô sẽ được phân tích bởi phương pháp phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA) để khái quát hóa dữ liệu một cách tổng quan, tìm ra những đặc điểm chính của dữ liệu, sau đó thông qua tiền xử lý và lấy mẫu, gán nhãn lần thứ nhất trước khi thực hiện các phương pháp máy học.

Dựa vào Hình 2, các biểu đồ hộp (Boxplot) cho thấy dữ liệu thô thu thập từ 04 ứng dụng thương mại di động có rất nhiều đánh giá có số lượng từ hoặc số lượng ký tự rất lớn (Hình 3), các giá trị ngoại lai này sẽ ảnh hưởng đến độ chính xác và gây nhiễu cho các phương pháp máy học.

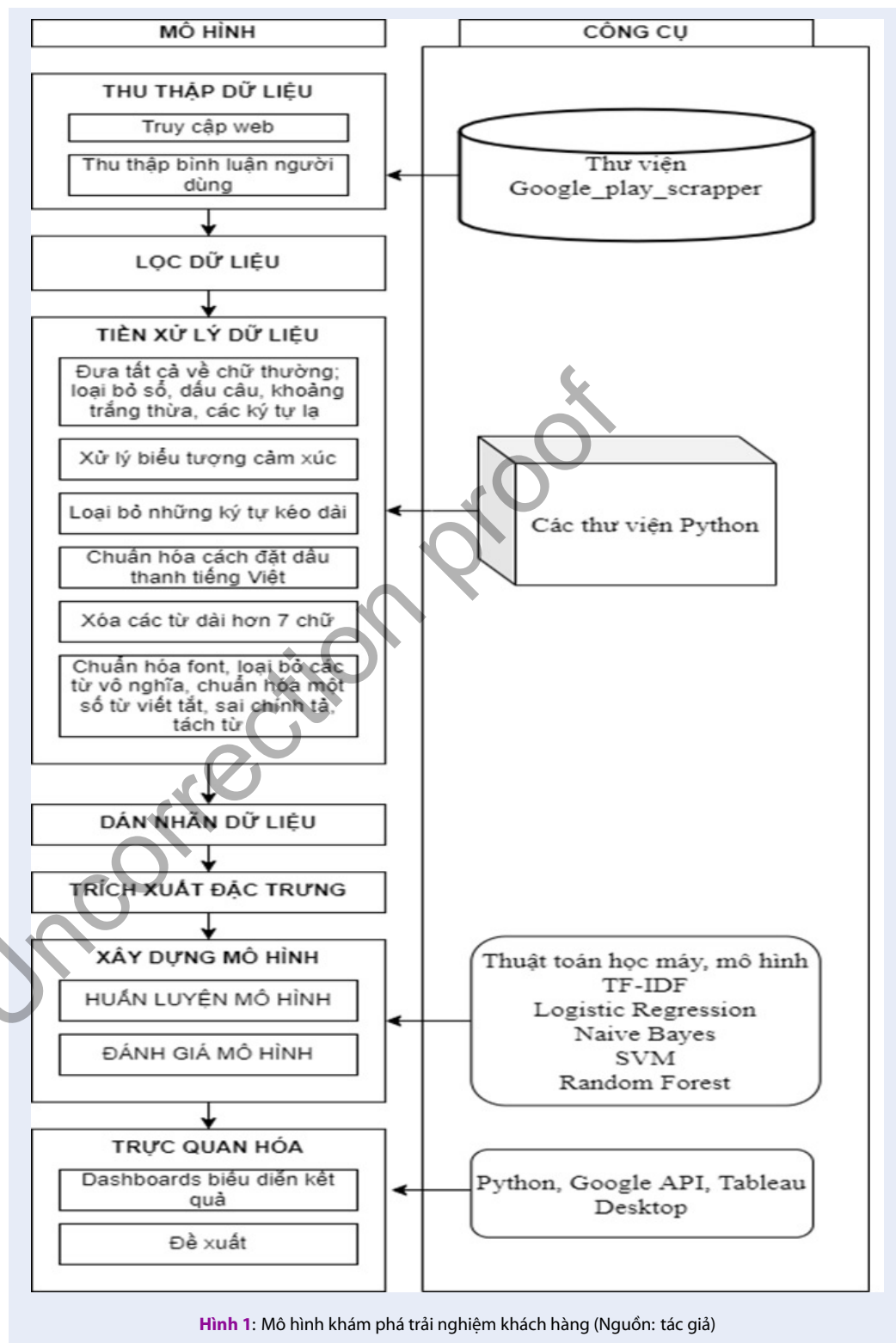
Vì vậy, chúng tôi đã lọc ra các giá trị ngoại lai nằm ngoài chặn trên (upper fence) và chặn dưới (lower fence) của các đánh giá thu thập được bằng cách loại bỏ các giá trị ngoại lai và thay thế các giá trị ngoại lai bằng IQR để cải thiện hiệu suất và độ tin cậy của các mô hình (kết quả được thể hiện trên Hình 4). Từ đó, đồng bộ được dữ liệu và tiến hành các bước xử lý tiếp theo.

Trong những nghiên cứu có liên quan đến ngôn ngữ tự nhiên, pha tiền xử lý dữ liệu vô cùng quan trọng và có ảnh hưởng đến kết quả phân tích. Khi thu thập bộ dữ liệu từ các ứng dụng thương mại di động điện tử, dữ liệu đang ở dạng thô (chưa được qua xử lý, có những trường bị rỗng, sai chính tả, chứa các ký tự đặc biệt hoặc các biểu tượng quan điểm,...), dữ liệu này sẽ làm giảm độ chính xác của kết quả các mô hình. Vì vậy, chúng tôi đã tiến hành các bước hiệu chỉnh góp phần tăng độ chính xác cho phân loại. Cuối cùng, tách từ là một phần không thể thiếu của quá trình xử lý ngôn ngữ tự nhiên. Chúng tôi đã sử dụng bộ công cụ mã nguồn mở xử lý ngôn ngữ tự nhiên Tiếng Việt Underthesea v1.3.4²⁶ để tách cụm từ, câu, đoạn văn, tài liệu văn bản thành các đơn vị nhỏ hơn (Tokens) để có thể dễ dàng diễn giải ý nghĩa của văn bản. Một số mẫu kết quả trước và sau khi thực hiện tiền xử lý được trình bày trong Bảng 2.

Sau khi thực hiện tiền xử lý, dữ liệu từ dữ liệu thô đã được chuẩn hóa đưa về cùng một dạng hỗ trợ cho việc huấn luyện và phân tích quan điểm khách hàng thông qua đánh giá để có được một kết quả dự đoán tốt nhất. Bảng 3 thể hiện số lượng các bình luận trước và sau khi tiền xử lý.

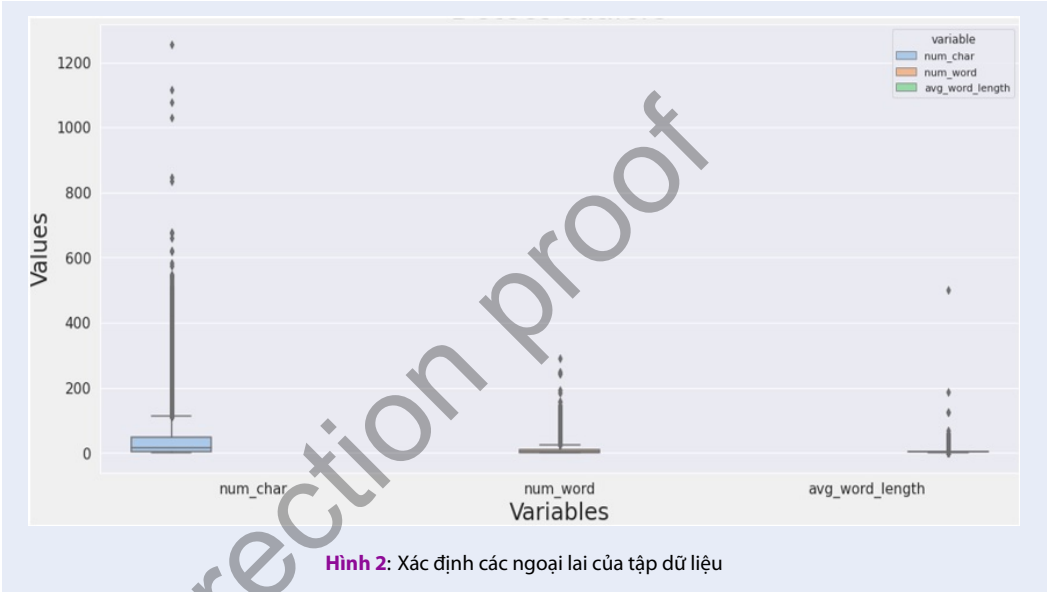
Dán nhãn dữ liệu

Bước này là để chuẩn bị một bộ dữ liệu được phân loại đủ lớn để được sử dụng như một bộ dữ liệu huấn luyện. Thông thường, các bộ dữ liệu này sẽ được xây dựng bằng tay để biết các nghiên cứu bằng các phương

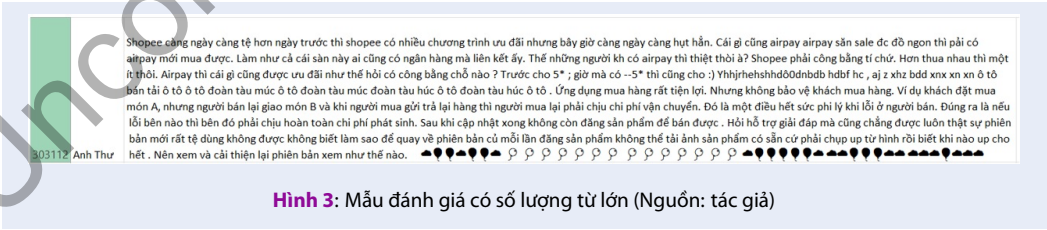


Bảng 1: Mẫu dữ liệu thu thập được từ các ứng dụng thương mại di động

userName	Content	at	score	thumbsUpCount	address
Lợi Lê thị	Uy tín chất lượng	01/12/2021	5	0	Shopee
Cường Trần	Ok	01/12/2021	5	0	Lazada
Bảo Nguyễn	tốt	01/12/2021	5	0	Lazada
Thư Ngọc	shopee siêu xịn luôn ạ!!	02/12/2021	5	10	Shopee
Hai Tieu	rất hài lòng .	03/12/2021	5	0	Tiki
Cong Tran	trang mua sắm tiện lợi	06/12/2021	5	0	Sendo



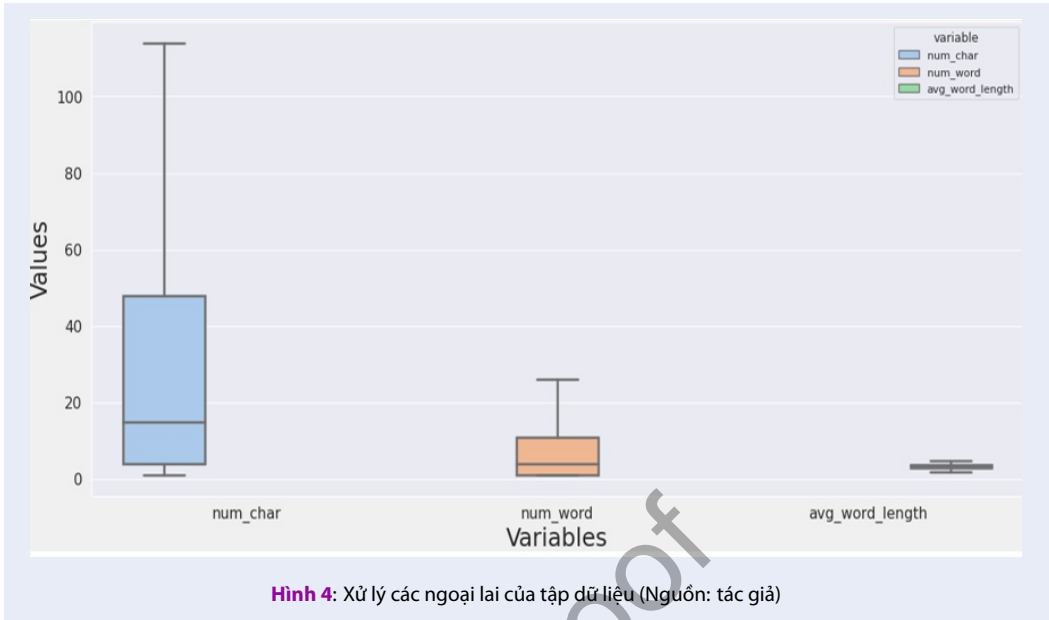
Hình 2: Xác định các ngoại lai của tập dữ liệu



Hình 3: Mẫu đánh giá có số lượng từ lớn (Nguồn: tác giả)

Bảng 2: Minh họa dữ liệu trước và sau khi tiến xử lý

Trước	Sau
Cho 5 sao ủng hộ chứ xài vẫn còn nhiều cái chán lắm:(nên cải thiện	cho sao_ủng_hộ chứ xài vẫn còn nhiều cái chán lắm nên cải thiện
Đúng nơi mua sắm cho mọi người	đúng nơi mua_sắm cho mọi người
Hài lòng ,	hài_lòng
Làm việc chậm chạp	làm_việc chậm_chạp
OK! Rất tốt! Rất hài lòng về cách phục vụ!	ổn rất tốt rất hài_lòng về cách phục_vụ
Phí ship quá đắt	phí_vận_chuyển quá đắt



Bảng 3: Số lượng các bình luận thu thập được trước và sau khi tiến xử lý

Ứng dụng	Số lượng trước	Số lượng sau
Tiki	67.698	57.788
Shopee	404.131	343.786
Lazada	459.981	386.062
Sendo	175.952	148.136
Tổng	1.107.762	935.772

pháp máy học. Tuy nhiên, trong nghiên cứu này, sau khi xem xét ngẫu nhiên các bộ dữ liệu được thu thập và đưa ra điểm đánh giá, nghiên cứu này tìm thấy các đánh giá với điểm đánh giá dưới 3,0 là tiêu cực, những điểm lớn hơn 3,0 là tích cực và trung tính khi điểm bằng 3,0. Do đó, bộ dữ liệu huấn luyện sẽ được phân loại và gán nhãn thành 2 lớp: tiêu cực, tích cực.

Phương pháp trích xuất đặc trưng

Để trích xuất đặc trưng từ tập dữ liệu chuẩn bị cho bước huấn luyện mô hình, trong các nghiên cứu của [27,28] đã đưa ra một số kỹ thuật tiền xử lý tập dữ liệu dạng văn bản, sau đó gán nhãn và sử dụng véc-tơ trọng số TF-IDF để đánh giá mức độ quan trọng của 1 từ và tần suất xuất hiện của từ đó trong đoạn văn bản, khi áp dụng kỹ thuật trích xuất đặc trưng này thì chúng ta sẽ xếp hạng được các véc-tơ đặc trưng cùng với các thuật toán phân cụm. Trong nghiên cứu này, sau bước tiền xử lý dữ liệu, phương pháp trích xuất đặc trưng TF-IDF [29,30] được áp dụng để xây dựng véc-tơ trọng số thể hiện tần suất xuất hiện của từ trong các

bình luận của khách hàng. Bảng 4 trình bày một số mẫu minh họa ma trận tần suất TF-IDF với các dòng thể hiện mỗi bình luận của khách hàng và các cột thể hiện trọng số của từ xuất hiện trong bình luận.

Thực nghiệm và đánh giá mô hình

Sau khi trích xuất đặc trưng toàn bộ tập dữ liệu với trên 935.000 bình luận, chúng tôi tiến hành huấn luyện phương pháp máy học cho tập dữ liệu huấn luyện. Dữ liệu được lấy mẫu được chia thành 2 nhóm: tập dữ liệu huấn luyện (80%) và tập dữ liệu kiểm tra (20%). Tập dữ liệu huấn luyện được sử dụng để thiết lập bởi phương pháp máy học bao gồm Hồi qui Logistic (LR) , Support Vector Machine (SVM), Naïve Bayes (NB) và Random Forest (RF) và sau đó áp dụng ma trận nhầm lẫn (Confusion Matrix) [31] với các độ đo Precision, Recall, F-score và Accuracy để đánh giá kết quả nhằm chọn ra mô hình phù hợp nhất, tiến hành gán nhãn lần thứ hai để áp dụng cho tập dữ liệu kiểm tra. Kết quả được thực nghiệm và đánh giá mô hình được thể hiện trên Bảng 5.

Bảng 4: Mẫu ma trận tần suất TF_IDF từ tập dữ liệu

	ổn	ứng_dụng	nhanch	ng nghiệ n	ki ể m_du ệ t	t ố t	t ố i_t ệ
Doc_1	0,4	0,0	0,3	0,0	0,2	0,1	0,0
Doc_2	0,0	0,2	0,0	0,7	0,0	0,0	0,1
Doc_3	0,2	0,0	0,2	0,4	0,0	0,5	0,0
Doc_4	0,0	0,0	0,5	0,0	0,4	0,2	0,0

Bảng 5: Kết quả đánh giá mô hình trên toàn bộ dữ liệu

Phương pháp máy học	Precision		Recall		F_score		Accuracy	Training time	Prediction time
	Pos	Neg	Pos	Neg	Pos	Neg			
LR	0,93	0,83	0,96	0,73	0,95	0,78	0,91	17,8s	1,3s
SVM	0,93	0,83	0,96	0,74	0,95	0,78	0,91	17h 38min 47s	28min 16s
NB	0,94	0,77	0,94	0,75	0,94	0,76	0,90	9,04s	1,33s
RF	0,94	0,82	0,96	0,76	0,95	0,79	0,92	41min 4s	24,3s

Từ bảng trên, chúng tôi nhận thấy rằng về độ chính xác thì thuật toán Random Forest là cao nhất (92%), lần lượt theo sau là Logistic Regression, Support Vector Machine (91%) và Naive Bayes (90%) là thấp nhất. Về thời gian huấn luyện mô hình thì thuật toán Naive Bayes là tốn ít thời gian nhất (9,04s), xếp theo sau là Logistic Regression (17,8s), Random Forest (41min 4s) và Support Vector Machine (17h 38min 47s). So về thời gian dự đoán, Logistic Regression là thuật toán sử dụng ít thời gian để dự đoán nhất (1,3s), tiếp theo là Naive Bayes (1,33s), Random Forest (24,3s) và Support Vector Machine (28min 16s). Với những chỉ số đánh giá trên, chúng tôi cho rằng với tập dữ liệu chúng tôi thu thập thì thuật toán Logistic Regression là phù hợp nhất.

Tập dữ liệu từ 04 ứng dụng thương mại di động (với Tiki gồm 57.788 bình luận, Shopee gồm 343.786 bình luận, Lazada gồm 386.062 bình luận và Sendo gồm 148.136 bình luận) sau khi được huấn luyện bằng phương pháp máy học dựa trên phân tích quan điểm đã cho kết quả khác nhau và thể hiện trên Bảng 6.

Sau khi thực nghiệm phương pháp máy học trên tập dữ liệu, kết quả cho thấy thuật toán SVM có thời gian huấn luyện và thời gian dự đoán lâu hơn tương đối nhiều so với các thuật toán khác, bởi vì bộ dữ liệu của 4 ứng dụng tương đối lớn nên mất nhiều thời gian để ánh xạ dữ liệu vào một không gian nhiều chiều hơn. Thời gian huấn luyện nhanh nhất Naive Bayes bởi vì

thuật toán này chạy dựa trên lý thuyết các biến dữ liệu độc lập với nhau, nhưng độ chính xác lại thấp hơn so với các thuật toán khác khi chạy trên các bộ dữ liệu của 04 ứng dụng thương mại di động thu thập được. Độ chính xác của thuật toán hồi quy Logistic cao hơn các thuật toán khác trong hầu hết các bộ dữ liệu (Tiki: 92%, Sendo: 90%, Shopee: 91%, and Lazada: 92%), chỉ thua độ chính xác của Random Forest trong bộ dữ liệu Lazada (93%). Tuy nhiên, thời gian huấn luyện và dự đoán lại nhanh hơn rất nhiều so với các thuật toán Random Forest và SVM. Kết quả cho thấy, Hồi quy Logistic là thuật toán tốt hơn so với các thuật toán khác khi xét về tổng thể thời gian dự đoán, huấn luyện cũng như là độ chính xác khi thực thi. Từ đó cho thấy rằng hồi quy Logistic phù hợp với bộ dữ liệu của 04 ứng dụng thương mại di động.

KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN

Nhìn vào các đánh giá được trực quan sau khi được áp dụng mô hình và xử lý, kết quả thực nghiệm đã cho thấy hiệu suất của việc phân loại dựa trên cảm tính từ tập dữ liệu về các đánh giá ứng dụng thương mại di động. Theo kết quả phân tích, tỷ lệ đánh giá tiêu cực (03 sao trở xuống) và tích cực (04 sao trở lên) trên 04 bộ dữ liệu của 04 ứng dụng thương mại di động tại Việt Nam cho thấy, tỷ lệ đánh giá tích cực của 04 ứng dụng đều ở mức cao (Tiki: 79,54%, Sendo: 83,99%, Shopee: 73,46%, Lazada: 83,80%).

Bảng 6: Kết quả so sánh các mô hình

Phương pháp máy học/Tập dữ liệu trên các ứng dụng		Precision		Recall		F_score		Accuracy	Training time	Prediction time
		Pos	Neg	Pos	Neg	Pos	Neg			
LR	Tiki	0,93	0,83	0,96	0,73	0,95	0,78	0,92	2,1s	104ms
	Sendo	0,92	0,76	0,97	0,55	0,94	0,63	0,90	2,93s	217ms
	Shopee	0,93	0,84	0,95	0,82	0,94	0,83	0,91	7,56s	603ms
	Lazada	0,94	0,82	0,97	0,67	0,96	0,74	0,92	6,88s	480ms
SVM	Tiki	0,93	0,82	0,96	0,73	0,95	0,77	0,91	1min	6,2s
	Sendo	0,92	0,76	0,97	0,54	0,94	0,63	0,90	15min 44s	34,5s
	Shopee	0,94	0,84	0,94	0,82	0,94	0,83	0,91	1h 21min 39s	4min 33s
	Lazada	0,94	0,82	0,97	0,67	0,96	0,74	0,92	2h 28min 28s	2min 21s
RF	Tiki	0,93	0,78	0,95	0,73	0,94	0,76	0,91	554ms	119ms
	Sendo	0,92	0,70	0,95	0,56	0,94	0,62	0,89	1,35s	218ms
	Shopee	0,94	0,80	0,93	0,83	0,93	0,81	0,90	3,67s	591ms
	Lazada	0,94	0,75	0,96	0,69	0,95	0,72	0,91	3,77s	478ms
	Tiki	0,94	0,81	0,95	0,76	0,95	0,78	0,92	32,6s	1,31s
	Sendo	0,92	0,74	0,96	0,55	0,94	0,63	0,90	2min 2s	2,97s
	Shopee	0,94	0,83	0,94	0,84	0,94	0,84	0,91	10min 39s	7,16s
	Lazada	0,94	0,82	0,97	0,70	0,96	0,75	0,93	7min 37s	8,07s

Biểu đồ thanh như trong Hình 5 được tạo ra để trực quan hóa những từ thường gặp nhất trong số tất cả các bình luận của khách hàng được phân tích. Biểu đồ đã cho thấy cách khách hàng nghĩ và cảm nhận về sản phẩm và dịch vụ trên các ứng dụng thương mại di động. Các từ “nghe”, “nhanh”, “tuyệt vời”, “tuyệt”,... là những từ thường gặp nhất trong các đánh giá mang tính tích cực, điều này cho thấy khách hàng có cảm nhận tốt đối với 04 ứng dụng thương mại di động được khảo sát. Từ “nhanh” và “nghe” là hai trong số những từ thường gặp nhất trong số tất cả các bình luận mang tính tích cực, cho thấy rằng hầu hết khách hàng đều có cảm nhận rất tích cực đối với bốn ứng dụng. Các từ “quảng cáo”, “tệ”, “chán”, “kém”,... là những từ thường gặp trong các bình luận tiêu cực, điều này cho thấy khách hàng vẫn có những cảm nhận chưa được tốt khi dùng các ứng dụng thương mại di động.

Hình 6 chỉ ra rằng hiệu suất của các đánh giá tích cực vượt trội hơn các đánh giá tiêu cực và trung bình chiếm hơn 75% trong tổng số các đánh giá thu thập được, từ đó, cho thấy số lượng đánh giá tích cực và mức độ hài lòng của khách hàng đối với các hoạt động trên các ứng dụng thương mại di động được khảo sát là cao. Điều này thể hiện sự hài lòng và tin tưởng của người dùng trong lĩnh vực thương mại di động. Nhìn vào ứng dụng Shopee, chúng ta dễ dàng nhận thấy tỷ lệ lượt đánh giá trong năm 2015 đến nay ngày càng tăng lên, chiếm hơn nửa tổng số các lượt đánh giá và đang dẫn đầu trong số các ứng dụng thương mại di động còn lại. Mặt khác, tỷ lệ lượt đánh giá của Sendo từ năm 2015 chiếm vị trí đứng đầu so với các ứng dụng thương mại di động khác nhưng đến năm 2021 đã giảm xuống hơn 8 lần so với trước và giữ vị trí thứ 3 chỉ đứng trước Tiki. Điều này cũng cho thấy được rằng, về tổng quan Shopee có sự phát triển mạnh mẽ nhất khi thu hút được thêm nhiều khách hàng qua số lượng bình luận được gia tăng hằng năm. Tiếp theo là về Lazada, khi số lượng bình luận, đánh giá có thể nói khá ổn định qua từng năm, tuy không phát triển nhanh như Shopee nhưng vẫn giữ vững được thị phần của mình và tiếp tục phát triển.

Dựa vào Hình 7 và Hình 8, các từ thể hiện quan điểm được sử dụng trong các đánh giá của khách hàng được hình ảnh hóa thông qua WordCloud. Từ đây, có thể thấy rằng các từ được khách hàng dùng nhiều trong các bình luận được thể hiện qua kích thước từ lớn để nhỏ dần. Cụ thể số lượng từ được dùng nhiều mang ý nghĩa tích cực (Hình 7) và tiêu cực (Hình 8), người quản lý có thể hình dung được khách hàng của mình đang quan tâm đến vấn đề gì. Chẳng hạn như từ “ngon” thì phần nhiều sẽ được hiểu là đang nói về thực phẩm/ thức ăn, điều này có nghĩa là thực phẩm



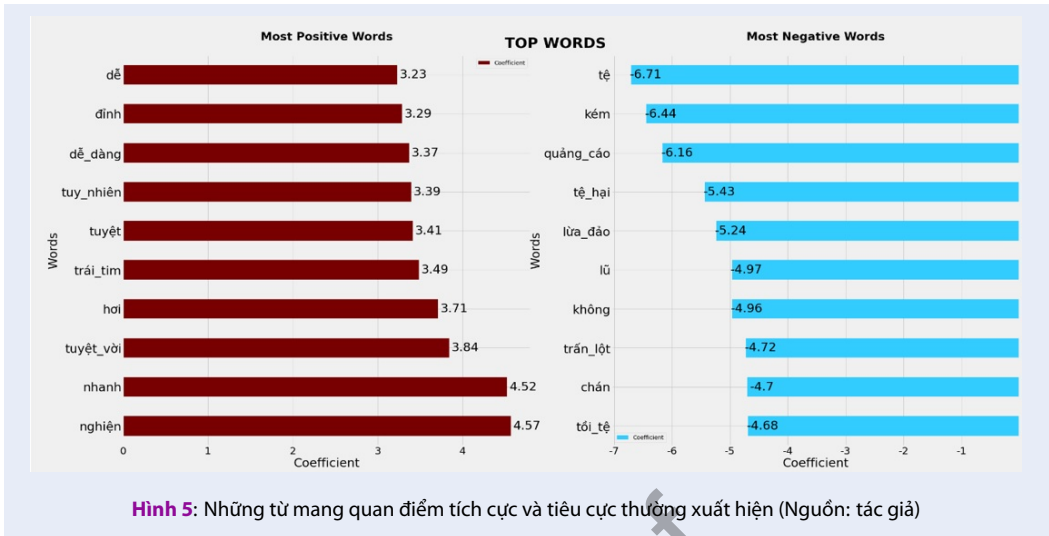
Hình 7: WordCloud từ Tích cực (Nguồn: tác giả)



Hình 8: WordCloud từ Tiêu cực (Nguồn: tác giả)

được bán trên ứng dụng được đánh giá tốt, còn với từ “đề”, có nghĩa là đề sử dụng các ứng dụng. Mặt khác, khi nói đến dịch vụ giao hàng, từ “nhanh” đại diện giao hàng hay thanh toán nhanh sau giao dịch. Những quan điểm hay khía cạnh mà khách hàng quan tâm sẽ giúp người quản lý nắm bắt tâm lý khách hàng một cách nhanh chóng và có chiến lược phát triển hiệu quả cho dịch vụ của mình.

Bên cạnh đó, xu hướng đánh giá chất lượng sản phẩm và dịch vụ của công ty cũng là cơ sở quan trọng để doanh nghiệp có thể có thêm cơ sở cải thiện chất lượng tốt hơn. Hình 8 đã chỉ ra kết quả về xu hướng đánh giá tiêu cực và tích cực theo từng giai đoạn thời gian, và thể hiện tổng quan về số lượng đánh giá tiêu cực và tích cực tại theo từng tháng từ năm 2015 đến năm 2021. Thông qua hai biểu đồ thể hiện trên Hình 8, ta có thể thấy năm 2015 đến 2017 thì thương



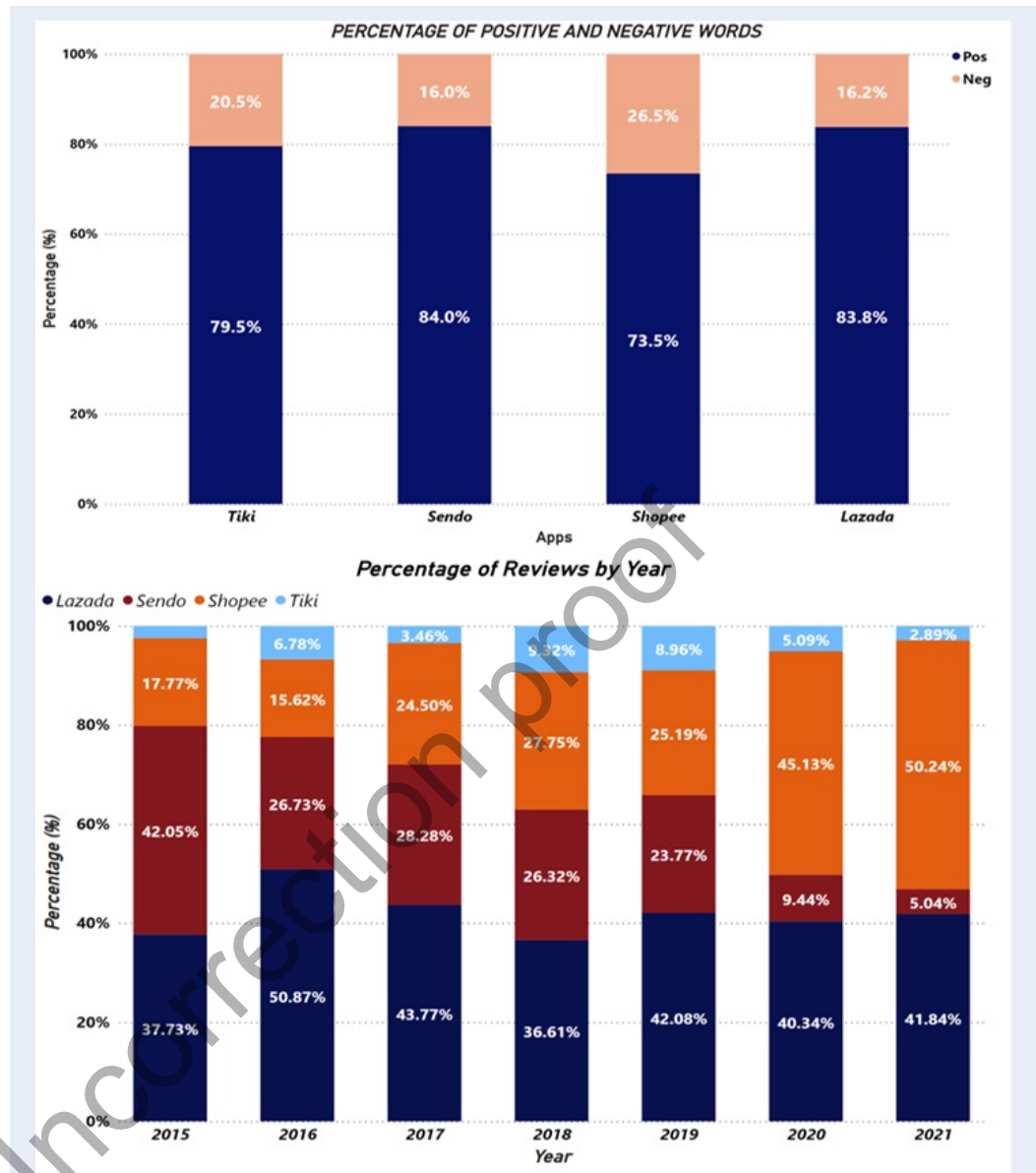
mại di động mới bước đầu xuất hiện tại Việt Nam nên số lượng đánh giá không thu thập được nhiều, nhưng sau đó số lượng đánh giá có sự phát triển theo thời gian. Từ năm 2018, thương mại di động tại Việt Nam tiếp tục phát triển toàn diện với mức tăng trưởng cao 30% so với các năm trước. Shopee là ứng dụng thương mại di động có mức độ hài lòng thấp với số lượng đánh giá tiêu cực liên tục gia tăng, đỉnh điểm là vào tháng 8 năm 2021 với khoảng 38% đánh giá tiêu cực, tiếp theo là Lazada, Sendo, Tiki. Shopee và Lazada là hai ứng dụng có số lượt đánh giá cao và cũng đang phát triển rất mạnh mẽ về mức độ nhận diện cũng như thương hiệu so với Sendo và Tiki.

Kết quả trên Hình 9 cũng cho thấy, tổng số lượng bình luận tích cực và tiêu cực từ năm 2019 đến năm 2021 bắt đầu tăng lên mạnh mẽ, ngay trong thời điểm đại dịch COVID-19 đang bùng phát (cuối năm 2019), khẳng định được nhu cầu mua sắm trực tuyến và việc sử dụng các ứng dụng thương mại di động ngày càng phổ biến như một xu hướng tất yếu khi so sánh với số lượng bình luận ở các năm trước. Đối với việc tăng lên hoặc giảm xuống của các bình luận mang tính chất tích cực, thể hiện cho sự hài lòng của khách hàng đối với các sản phẩm và dịch vụ mà họ trải nghiệm trên 4 ứng dụng thương mại di động, kết hợp với các biểu đồ trực quan khác (như Hình 5 và Hình 7) để thấy rõ vấn đề mà khách hàng có nhận xét tích cực, từ đó doanh nghiệp có thể đánh giá tổng quan chiến lược kinh doanh đang áp dụng có phù hợp với nhu cầu của người tiêu dùng hay không dựa vào quan điểm của họ đối với các sản phẩm và dịch vụ của doanh nghiệp. Thêm vào đó, cũng như các bình luận tích cực thì với các bình luận tiêu cực, doanh nghiệp có thể tìm ra được lỗ hổng trong các dịch vụ của mình, từ đó

cải thiện lại các hoạt động nhằm đem lại trải nghiệm tốt cho khách hàng của mình, điều này giúp doanh nghiệp nhìn vào dữ liệu mà khách hàng để lại trực tiếp từ ứng dụng (lấy dữ liệu tức thời và xử lý) thay vì phải chủ động đi khảo sát thị trường từng nhóm khách hàng, điều này làm mất nhiều thời gian và kém hiệu quả trong việc bắt kịp xu hướng mua sắm của

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Sau quá trình nghiên cứu này, mô hình phân tích quan điểm được đề xuất và thực nghiệm trên bộ dữ liệu trong lĩnh vực thương mại di động và kết quả đánh giá mô hình với tính chính xác cao. Việc tiến hành thực nghiệm và so sánh 04 phương pháp máy học (Naive Bayes, SVM, Hồi quy Logistic và Random Forest) đã phản ánh những đặc điểm, thuộc tính của từng thuật toán đối với bộ dữ liệu nghiên cứu thông qua các chỉ số của ma trận nhầm lẫn như Accuracy, F_Score. Các kết quả từ việc phân loại quan điểm (tích cực và tiêu cực) qua bình luận khách hàng đã cho thấy nhiều thông tin hữu ích về quan điểm và hành vi khách hàng, giúp các doanh nghiệp xác định được nhu cầu khách hàng và đưa ra gợi ý sản phẩm phù hợp cho những khách hàng tiềm năng. Qua những báo cáo, biểu đồ trực quan hóa dữ liệu, các doanh nghiệp có thể phân tích những mong muốn của khách hàng về nhiều khía cạnh trên các ứng dụng thương mại di động (chất lượng sản phẩm, đề xuất tìm kiếm, phản hồi, dịch vụ vận chuyển, ưu đãi giảm giá, vv...), từ đó đưa ra những chiến lược tối ưu nâng cao trải nghiệm khách hàng và tăng lợi thế cạnh tranh của thương hiệu. Bên cạnh đó, mô hình của nghiên cứu có thể được tích hợp thêm vào ứng dụng nhằm khảo

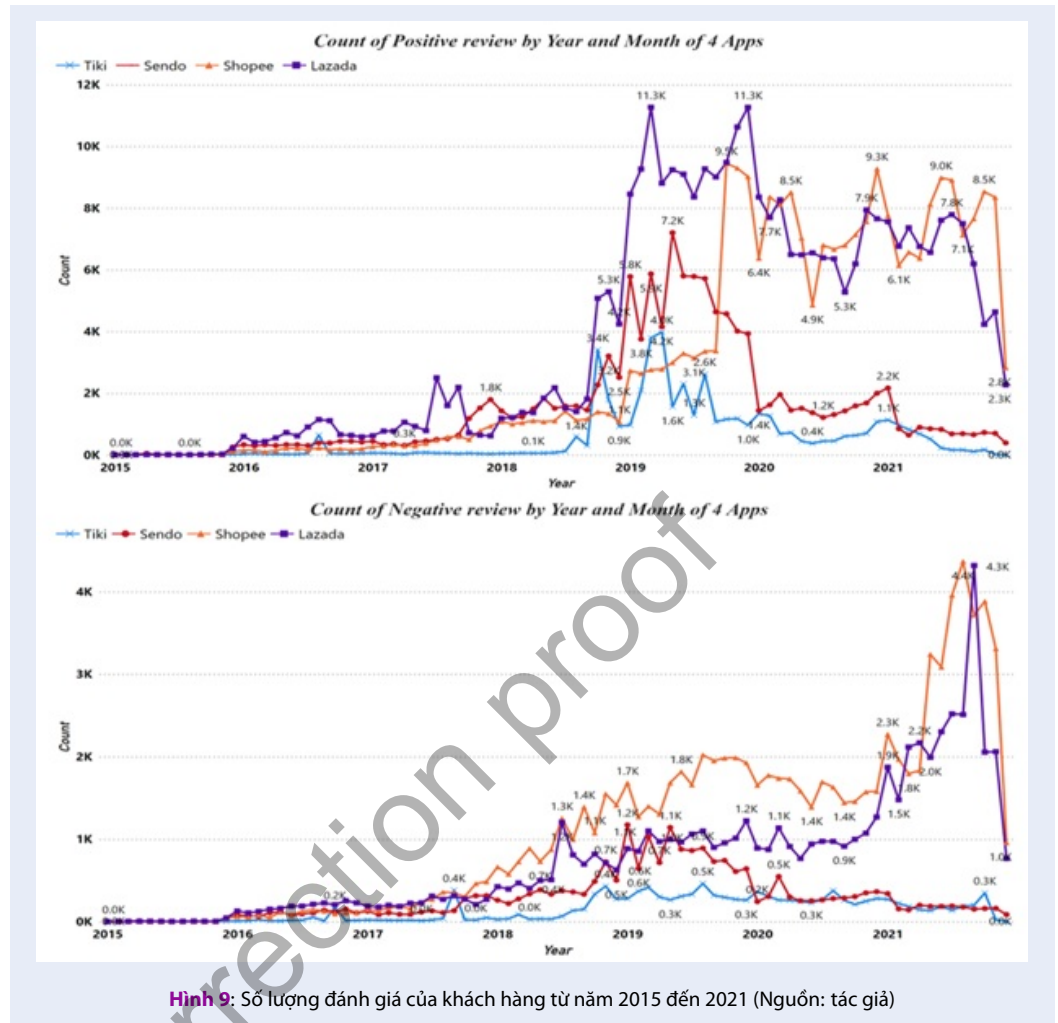


Hình 6: Phân bố phần trăm đánh giá theo ứng dụng và theo năm (Nguồn: tác giả)

sát quan điểm khách hàng đối với nhiều lĩnh vực, dịch vụ và sản phẩm khác nhau của các doanh nghiệp. Bộ dữ liệu thực nghiệm được thu thập trong khoảng thời gian từ năm 2015 đến năm 2021, đảm bảo tính thực tế và phản ánh được sự biến động trong thời gian dài. Điều này giúp các nhà phân tích tìm ra được xu hướng thị trường và đưa ra các dự đoán cho tương lai, từ đó có những chiến lược đầu tư hợp lý giảm thiểu rủi ro ở mức thấp nhất. Có thể thấy, việc ứng dụng phân tích quan điểm qua những bình luận bằng các phương pháp máy học và xử lý ngôn ngữ tự nhiên là giải pháp cần thiết nhằm nâng cao chất lượng sản phẩm, dịch

vụ của doanh nghiệp.

Trong tương lai, để cải thiện hiệu suất và tính chính xác của mô hình, phương pháp BERT trong xử lý ngôn ngữ tự nhiên, trích xuất các khía cạnh với kỹ thuật gắn thẻ từng phần (POS Tagging), phân tích các biểu tượng cảm xúc trong các bình luận sẽ được nghiên cứu và khai thác. Bên cạnh đó, nghiên cứu tiếp tục tối ưu hóa mô hình để xây dựng thành một hệ thống đánh giá, phân loại bình luận người dùng với cơ chế hoạt động liên tục thu thập dữ liệu, áp dụng mô hình phân loại đưa ra các báo cáo trực quan hỗ trợ doanh nghiệp ra quyết định; mở rộng mô hình nghiên cứu



sang các lĩnh vực khác không chỉ dừng lại ở thương mại đi động, khai thác mọi tương tác và bình luận của khách hàng trực tuyến với mục đích mang lại những lợi ích tối đa cho doanh nghiệp và người dùng.

DANH MỤC TỪ VIẾT TẮT

- EDA: Exploratory Data Analysis
- TF_IDF: Term Frequency – Inverse Document Frequency
- LR: Logistic Regression
- SVM: Support Vector Machine
- NB: Naive Bayes
- RF: Random Forest
- BERT: Bidirectional Encoder Representations from Transformers
- Pos: Positive
- Neg: Negative
- IQR: Interquartile Range

XUNG ĐỘT LỢI ÍCH

Nhóm tác giả xin cam đoan rằng không có bất kỳ xung đột lợi ích nào trong công bố bài báo.

ĐÓNG GÓP TÁC GIẢ

Toàn bộ nội dung bài báo do nhóm tác giả thực hiện. Đóng góp của từng tác giả với nội dung bài báo:

Nguyễn Trần Thúy Quỳnh, Bùi Nguyễn Bích Ngọc, Nguyễn Thị Bảo Trâm, Hồ Trung Thành chịu trách nhiệm nội dung: khảo sát các nghiên cứu liên quan, xây dựng mục tiêu, đề xuất phương pháp và mô hình nghiên cứu, đánh giá kết quả và thảo luận.

Trần Nhật Nguyên, Võ Bá Tùng và Hồ Trung Thành chịu trách nhiệm nội dung: Khảo sát các nghiên cứu liên quan, nghiên cứu về ý tưởng, triển khai ý tưởng dựa trên phương pháp nghiên cứu, phân tích dữ liệu, và trực quan kết quả phân tích.

534 TÀI LIỆU THAM KHẢO

- 535 1. The value and role of data in electronic commerce and the dig-
536 ital economy and its implications for inclusive trade and de-
537 velopment [internet]. UNCTAD; 2019;Available from: [https://](https://unctad.org/system/files/official-document/tdb_edc3d2_en.pdf)
538 unctad.org/system/files/official-document/tdb_edc3d2_en.pdf.
- 539 2. Ritter T, Pedersen CL. Digitization capability and the digital-
540 ization of business models in business-to-business firms: past,
541 present, and future. *Ind Mark Manag*. 2020;86:180-90;
- 542 3. Liu B. Sentiment analysis and opinion mining. San Rafael, CA:
543 Morgan & Claypool; 2012;Available from: [https://doi.org/10.](https://doi.org/10.1007/978-3-031-02145-9)
544 [1007/978-3-031-02145-9](https://doi.org/10.1007/978-3-031-02145-9).
- 545 4. Sharma R, Nigam S, Jain R. Opinion mining in Hindi language:
546 A survey. *UFCST*. 2014;4(2):41-7;Available from: [https://doi.](https://doi.org/10.5121/ijfcs.2014.4205)
547 [org/10.5121/ijfcs.2014.4205](https://doi.org/10.5121/ijfcs.2014.4205).
- 548 5. Bang B, Lee L. Opinion mining and sentiment analysis. *Found*
549 *Trends Inf Retrieval*. 2008;2(1-2):1-135;Available from: [https://](https://doi.org/10.1561/15000000011)
550 doi.org/10.1561/15000000011.
- 551 6. Wanner F, Rohrdant C, Mansmann F, Oelke D, Keim D. Visual
552 sentiment analysis of RSS News Feeds featuring the US presi-
553 dential election in 2008 (VISSW 2009); 2009. Visual Interfaces
554 to the Social and the Semantic [web].;
- 555 7. Nguyen B, Nguyen V-H, Ho T. Sentiment analysis of customer
556 feedback in online food ordering services. *Bus Syst Res J*.
557 2021;12(2):46-59;Available from: [https://doi.org/10.2478/bsrj-](https://doi.org/10.2478/bsrj-2021-0018)
558 [2021-0018](https://doi.org/10.2478/bsrj-2021-0018).
- 559 8. Binali H, Potdar V, Wu C. A state of the art opinion mining
560 and its application domains IEEE International Conference on
561 Industrial Technology. Vol. 2009; 2009;Available from: [https://](https://doi.org/10.1109/ICIT.2009.4939640)
562 doi.org/10.1109/ICIT.2009.4939640.
- 563 9. Ruder S, Ghaffari P, Breslin G J. A Hierarchical Model of Reviews
564 for Aspect-based Sentiment Analysis. *EMNLP*. 2016;7;Available
565 from: <https://doi.org/10.18653/v1/D16-1103>.
- 566 10. Al-Otaibi S, Alnassar A, Alshahrani A, Al-Mubarak A, Albugami
567 S, Almutiri N et al. Customer satisfaction measurement using
568 sentiment analysis. *Int J Adv Comput Sci Appl*. 2018;9(2);Avail-
569 able from: <https://doi.org/10.14569/IJACSA.2018.090216>.
- 570 11. Mudinas A, Zhang D, Levene M. Combining lexi-
571 con and learning based approaches for concept-level
572 sentiment analysis. *Proceedings of the first interna-*
573 *tional workshop on issues of sentiment discovery*
574 *and opinion mining. Wisdom*. 2012;12;Available from:
575 <https://doi.org/10.1145/2346676.2346681>.
- 576 12. Pang B, Lee L, Vaithyanathan S. Thumbs up? *Proceedings*
577 *of the ACL-02 conference on Empirical methods in natural*
578 *language processing - EMNLP*. Vol. 2002; '02;Available from:
579 <https://doi.org/10.3115/1118693.1118704>.
- 580 13. Vu L, Le T. A lexicon-based method for Sentiment Analysis
581 using social network data. *Information and Knowledge Engi-*
582 *neering (IKE)*. 2017;10-6;
- 583 14. Taj S, Shaikh B, Fatemah Meghji A. Sentiment Analysis of News
584 Articles: A Lexicon based Approach 2nd International Confer-
585 ence on Computing, Mathematics and Engineering Technolo-
586 gies (iCoMET). Vol. 2019; 2019;Available from: [https://doi.org/](https://doi.org/10.1109/ICOMET.2019.8673428)
587 [10.1109/ICOMET.2019.8673428](https://doi.org/10.1109/ICOMET.2019.8673428).
- 588 15. Ding X, Liu B, Yu P. A holistic lexicon-based approach to opin-
589 ion mining. *Proceedings of the international conference on*
590 *web search and web data mining - WSDM*. Vol. 2008; '08;Avail-
591 able from: <https://doi.org/10.1145/1341531.1341561>.
- 592 16. Palanisamy P, Yadav V, Elchuri H. Serendio: simple and Practi-
593 cal lexicon based approach to Sentiment Analysis; 2013;
- 594 17. Poria S, Chaturvedi I, Cambria E, Bisio F. Sentic LDA: improv-
595 ing on LDA with semantic similarity for aspect-based senti-
596 ment analysis. 2016 International Joint Conference on Neu-
597 ral Networks (IJCNN);2016;Available from: [https://doi.org/10.](https://doi.org/10.1109/IJCNN.2016.7727784)
598 [1109/IJCNN.2016.7727784](https://doi.org/10.1109/IJCNN.2016.7727784).
- 599 18. Ruder S, Ghaffari P, Breslin G J. A Hierarchical Model of Reviews
600 for Aspect-based. Sentiment analysis. *EMNLP*. 2016;7;Avail-
601 able from: <https://doi.org/10.18653/v1/D16-1103>.
- 602 19. Hutto C, VADER GE. A parsimonious rule-based model for sen-
603 timent analysis of social media text. *ICWSM*. 2015;
- 604 20. Domingos P, Pazzani M. *Mach Learn*. 1997;29(2/3):103-30. doi:
605 10.1023/A:1007413511361;Available from: [https://doi.org/10.](https://doi.org/10.1023/A:1007413511361)
606 [1023/A:1007413511361](https://doi.org/10.1023/A:1007413511361).
- 607 21. Maalouf M. Logistic regression in data analysis: an overview.
608 *Int J Data Anal Tech Strateg*. 2011;3(3):281;Available from:
609 <https://doi.org/10.1504/IJDATS.2011.041335>.
- 610 22. Cutler A, Cutler D, Stevens J. Random forests. *Ensemble Mach*
611 *Learn*. 2012;157-75;Available from: [https://doi.org/10.1007/](https://doi.org/10.1007/978-1-4419-9326-7_5)
612 [978-1-4419-9326-7_5](https://doi.org/10.1007/978-1-4419-9326-7_5).
- 613 23. Neshan S, Akbari R. A combination of machine
614 learning and lexicon based techniques for senti-
615 ment analysis 6th International Conference on Web
616 Research (ICWR). Vol. 2020; 2020;Available from:
617 <https://doi.org/10.1109/ICWR49608.2020.9122298>.
- 618 24. Dang NC, Moreno-García MN, De la Prieta F. Sentiment anal-
619 ysis based on deep learning: A comparative study. *Elec-*
620 *tronics*. 2020;9(3):483;Available from: [https://doi.org/10.3390/](https://doi.org/10.3390/electronics9030483)
621 [electronics9030483](https://doi.org/10.3390/electronics9030483).
- 622 25. Nandi A, Sharma P. Comparative study of sentiment anal-
623 ysis techniques. *Interdiscip Res Technol Manag*. 2021:456-
624 60;Available from: <https://doi.org/10.1201/9781003202240-72>.
- 625 26. GitHub - undertheseanlp/underthesea: Underthesea - Viet-
626 namese NLP Toolkit; 2022 [cited May 20 2022];Available from:
627 [https://github.com/undertheseanlp/underthesea#2-word-](https://github.com/undertheseanlp/underthesea#2-word-segmentation)
628 [segmentation](https://github.com/undertheseanlp/underthesea#2-word-segmentation).
- 629 27. Ahmed H, Awan M, Khan N, Yasin A, Faisal Shehzad H. Senti-
630 ment analysis of online food reviews using big data analytics.
631 *Elem Educ Online*. 2021;20(2):827-36;
- 632 28. Nasim Z, Rajput Q, Haider S. Sentiment analysis of student
633 feedback using machine learning and lexicon based ap-
634 proaches International Conference on Research and Innova-
635 tion in Information Systems (ICRIIS). Vol. 2017; 2017;Available
636 from: <https://doi.org/10.1109/ICRIIS.2017.8002475>.
- 637 29. Rajaraman A[and Ullman, J., n.d. Data mining. Mining of Mas-
638 sive Datasets, pp. 1-17;.
- 639 30. Beel J, Gipp B, Langer S, Breiteringer C. Research-paper
640 recommender systems: a literature survey. *Int J Digit*
641 *Libr*. 2016;17(4):305-38;Available from: [https://doi.org/10.](https://doi.org/10.1007/s00799-015-0156-0)
642 [1007/s00799-015-0156-0](https://doi.org/10.1007/s00799-015-0156-0).
- 643 31. Visa S, Ramsay B, Ralescu A, Knaap E. Confusion matrix-
644 based feature selection. *CiteSeerX*; 2022;Available from: [http:](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.666.8961)
645 [/citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.666.8961](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.666.8961).

Customer experience discovery model based on sentiment analysis and machine learning method

Quynh Nguyen Tran Thuy^{1,2}, Ngoc Bui Nguyen Bich^{1,2}, Tram Nguyen Thi Bao^{1,2}, Nguyen Tran Nhat^{1,2}, Tung Ba Vo^{1,2}, Thanh Ho Trung^{1,2,*}



Use your smartphone to scan this QR code and download this article

ABSTRACT

Exploiting customer insights plays a key role in the long-term development strategy of every business, which strongly supports them in building trust with customers and enhancing competitive advantages. Especially in the period of rapid growth of technology, as there is a significant increase in online users, all the interactions and comments become more important in extracting valuable insights into customer behaviours and emotions. Our research aims to conduct sentiment analysis on comments in mobile commerce apps, then classify them into positive or negative sentiment, based on a hybrid approach that combines supervised machine learning methods and natural language processing techniques; at the same time, evaluating the model performance through the Confusion Matrix to ensure the reliability of the results. The study experimented on a proposed model with 04 machine learning methods including Naive Bayes, SVM, Random Forest and Logistic Regression on more than 935,000 comments collected from 04 popular mobile commerce apps in Vietnam (Tiki, Shopee, Lazada and Sendo). The experiment categorised positive and negative views with high accuracies of 91%, expressed by reports and charts that reflect customer trends and feelings. Moreover, the study also brings new and deeper perspectives on customer behaviours, assisting administrators to detect the strengths and weaknesses of services and apps, thereby improving user experience. Based on the research results, E-commerce businesses can analyse market trends and explore customers' needs and interests to develop effective product and service development strategies.

Key words: Sentiment analysis, machine learning, classification models, customer reviews, mobile electronic commerce

¹University of Economics and Law, Ho Chi Minh City, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

Correspondence

Thanh Ho Trung, University of Economics and Law, Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City, Vietnam

Email: thanhht@uel.edu.vn

History

- Received: 20-3-2022
- Accepted: 09-9-2022
- Published:

DOI :



Copyright

© VNU-HCM Press. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.



Cite this article : Thuy Q N T, Bich N B N, Bao T N T, Nhat N T, Vo T B, Trung T H. **Customer experience discovery model based on sentiment analysis and machine learning method.** *Sci. Tech. Dev. J. - Eco. Law Manag.*; (1):1-1.