Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế-Luật và Quản lý 2022, 6(4):3682-3690

Open Access Full Text Article

Bài nghiên cứu

Ứng dụng phương pháp học máy để phân tích bình luận của khách hàng về thực phẩm tươi sống trên các sàn thương mại điện tử ở Việt Nam

Nguyễn Thị Ngọc Ánh, Phan Thị Hà Giang, Võ Chí Giang, Nguyễn Bá Thịnh An, Nguyễn Phát Đạt, Hồ Thy Nhân Ái, Nguyễn Quang Hưng*

TÓM TẮT

Trong những năm gần đây, các hộ nông dân đã phát triển việc bán nông sản cho người tiêu dùng thông qua các sàn TMĐT. Rõ ràng là TMĐT đã trở thành một phương thức mới và hiệu quả để giúp các hộ nông dân tiếp cận thị trường. So với các mặt hàng khác, nông sản là mặt hàng bị ảnh hưởng nặng nề theo mùa vụ, có những đặc điểm khó khăn như thời hạn sử dụng ngắn, dễ bị hư hỏng, chi phí bảo quản và vận chuyển cao. Người tiêu dùng đặt ra những tiêu chuẩn cao về chất lượng, tốc độ giao hàng, tần suất tiêu thụ, đơn giá cho các sản phẩm này. Việc lấy đánh giá, phản ánh của khách hàng làm đối tượng nghiên cứu giúp doanh nghiệp khám phá cơ chế ra quyết định của người tiêu dùng, từ đó có chiến lược tiếp thị phù hợp cho nông sản của mình. Bên cạnh đó, họ sẽ thấy được khách hàng không hài lòng với điểm nào để khắc phục, nâng cao chất lượng sản phẩm và dịch vụ. Trong bài nghiên cứu này, nhóm tác giả nghiên cứu và đề xuất các phương pháp học máy để phân loại, sàng lọc và khai phá bình luận dạng văn bản của khách hàng về các sản phẩm nông sản trên 3 sàn TMĐT Tiki, Sendo và Voso. Thực nghiệm mô hình trên tập dữ liệu thu thập được với kết quả thuật toán sgdclassifier kết hợp với phương pháp One-vs-Rest cho kết quả dự đoán tốt nhất với 87%. Nghiên cứu cũng xây dựng các biểu đồ và thể hiện trực quan số liệu phân tích những yếu tố tác động đến sự hài lòng của khách hàng về chất lượng sản phẩm cũng như dịch vụ của người bán và sàn TMĐT. Ngoài ra, nghiên cứu đề xuất những khuyến nghị giúp cho doanh nghiệp có thể cải thiện chất lượng sản phẩm, dịch vụ từ đó có chiến lược để thu hút và giữ chân khách hàng tốt hơn.

Từ khoá: bình luận người dùng, học máy, nông sản, phân tích cảm xúc, thương mại điện tử

Trường Đại học Kinh tế - Luật, ĐHQG-HCM

Liên hệ

Nguyễn Quang Hưng, Trường Đại học Kinh tế - Luật, ĐHQG-HCM

Email: hungnq@uel.edu.vn

Lịch sử

- Ngày nhận: 18-9-2022
- Ngày chấp nhận: 16-12-2022
- Ngày đăng: 31-1-2023

DOI :

https://doi.org/10.32508/stdjelm.v6i4.1132

GIỚI THIỆU

TMĐT đang ngày càng phát triển và chiếm vai trò đáng kể trong nền kinh tế. Nông nghiệp từ xưa đến nay luôn là một ngành chủ lực của Việt Nam. Việc kết hợp cả hai ngành chắc chắn sẽ giúp kinh tế Việt Nam phát triển theo hướng tích cực. Các doanh nghiệp tham gia vào thị trường TMĐT trong lĩnh vực nông sản ngày càng nhiều vì thế việc cạnh tranh là điều bắt buộc. Để có thể nâng cao hiệu quả cạnh tranh, các doanh nghiệp cần hiểu rõ nhu cầu của khách hàng thông qua các bình luận của khách hàng. Để giải quyết bài toán này, nghiên cứu đã thu thập các bình luận về nông sản trên các sàn TMĐT. Nhưng dữ liệu chỉ ở mức sơ cấp, do đó, các phương pháp học máy đã được áp dụng vào nghiên cứu để có thể phân loại ra được các bình luận tích cực hay tiêu cực và kết hợp với các nhãn chủ đề liên quan như chất lượng nông sản, giá cả, dịch vụ, giao hàng, hệ thống trực tuyến và dùng các phương pháp phân tích, trực quan hóa dữ liệu trên các biểu đồ. Bài nghiên cứu sẽ đưa ra cái nhìn tổng quan về sản phẩm nông sản trên các sàn TMĐT (cụ thể ở đây là Tiki, Sendo, Voso) và các ngành hàng (rau, củ, trái cây, thịt), bên cạnh đó là đề xuất một mô hình phân tích cảm xúc dựa trên bình luận của họ về các sản phẩm nông sản trên các sàn TMĐT. Các biểu đồ được phân tích sẽ chỉ ra những yếu tố tác động đến sự hài lòng của khách hàng về chất lượng sản phẩm cũng như dịch vụ của người bán và sàn TMĐT.

CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

Khai phá văn bản, còn được gọi là phân tích văn bản, là một kỹ thuật trí tuệ nhân tạo để chuyển đổi dữ liệu phi cấu trúc thành dữ liệu có cấu trúc bằng cách sử dụng NLP phân tích bằng các thuật toán học máy [1]. Các kỹ thuật khai thác văn bản như phân tích cấp độ từ (ví dụ: phân tích tần suất), phân tích liên kết từ (ví dụ: network analysis) và các kỹ thuật nâng cao (ví dụ: phân loại văn bản, phân cụm văn bản, mô hình hóa chủ đề, truy xuất thông tin và phân tích cảm xúc) [2].

Trích dẫn bài báo này: Ánh N T N, Giang P T H, Giang V C, An N B T, Đạt N P, Ái H T N, Hưng N Q. **Ứng dụng phương pháp học máy để phân tích bình luận của khách hàng về thực phẩm tươi sống trên các sàn thương mại điện tử ở Việt Nam**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 6(4):3682-3690.

Trong bài nghiên cứu này, nhóm tác giả tập trung vào việc phân tích cảm xúc thông qua bình luận khách hàng.

Các kỹ thuật phân tích cảm xúc có thể chia thành hai loại là cách tiếp cận dựa trên từ vựng và cách tiếp cận dựa trên máy học [3]. Ngoài ra, có một phương pháp kết hợp kết hợp một số thuật toán phân loại cơ sở để tối ưu kết quả phân loại cuối cùng được gọi là các phương pháp tổng hợp (ensemble methods).

Hassan et.al [4] tiến hành nghiên cứu để cải thiện mô hình phân tích cảm xúc của Twitter. Nghiên cứu đã so sánh kết quả giữa các thuật toán: SVM, Logistic Regression, Naïve Bayes, Bayes Net, REP Tree, Random Tree, và RBF Neural Network. Kết quả cho thấy phương pháp tổng hợp đạt được độ chính xác cao hơn. Wang et.al [5] cũng tiến hành phân loại cảm xúc dùng 5 thuật toán là Support Vector Machine, K Nearest Neighbor, Decision Tree, Maximum Entropy và Naive Bayes. Thêm vào đó, nghiên cứu áp dụng thêm 3 phương pháp tổng hợp là Random Subspace, Boosting, and Bagging. Kết quả cho thấy Random Subspace cho ra kết quả tốt nhất.

Các nghiên cứu trước đó đã đạt được những kết quả tốt trong lĩnh vực phân tích cảm xúc. Trong nghiên cứu này, hóm tiến hành xây dựng mô hình tổng hợp 2 phương pháp Binary Relevance và One-vs-Rest. Mỗi phương pháp sẽ kết hợp với một thuật toán học máy để tiến hành huấn luyện và dự đoán dựa trên tập kiểm thử đã được chia trước đó. Điều này sẽ tối ưu hóa kết quả dự đoán bình luận khách hàng, phân tích từng cụm từ để xác định cảm xúc tích cực hay tiêu cự.

PHƯƠNG PHÁP NGHIÊN CỨU

Hình 1 trình bày mô hình nghiên cứu bình luận của khách hàng về nông sản trên các sàn thương mại điện tử dựa trên phương pháp học máy. Mô hình bài toán được chia ra làm 4 phần: Thu thập dữ liệu, Tiền xử lý dữ liệu, Huấn luyện mô hình và Phân tích trực quan hóa. Sử dụng các thư viện Request và BeautifulSoup trong ngôn ngữ lập trình Python để thu thập dữ liệu từ các sàn TMĐT như Tiki, Sendo và Voso. Các dữ liệu đầu vào được xử lý sạch như xóa dòng rỗng, kí tự icon, kí tự đặc biệt, chuyển về kí tự thường,... trước khi được đưa vào huấn luyện mô hình thông qua các thư viện có sẵn trong ngôn ngữ lập trình Python, việc này làm cho dữ liệu thô được điều chỉnh lại phù hợp với các bước sau. Dữ liệu còn được hệ thống hóa và gán nhãn dựa vào những khía cạnh khác nhau, sự quan tâm hay thái độ,... để phục vụ cho việc dự đoán. Các thuật toán học máy được sử dụng để huấn luyện kết hợp với các ensemble method để đưa ra độ chính xác cao nhất. Cuối cùng là thực hiện trực quan hóa dữ liệu với Power BI thông qua các biểu đồ từ đó có cơ sở đưa ra những đề xuất giải pháp.

Thu thập dữ liệu

Dựa trên lập trình bằng ngôn ngữ Python, chúng tôi sử dụng một số thư viện sẵn có để giúp cho việc khai phá dữ liệu như: Request, Pandas, NumPy. Tùy vào mỗi cấu trúc của website mà quyết định thu thập thông tin qua cấu trúc Hypertext Markup Language (HTML) hay Application Programming Interface (API) của website.

Mô tả dữ liệu

Bộ dữ liệu được nhóm nghiên cứu sử dụng gồm có dữ liệu từ 3 sàn giao dịch TMĐT là Tiki, Sendo và Voso với chủ đề về nông sản Việt Nam từ tháng 10/2021 trở về trước và từ sau tháng 10/2021 đến đầu tháng 05/2022. Chi tiết là các sản phẩm về rau củ quả, thịt, trái cây của sàn Tiki, các sản phẩm thịt trứng, rau củ quả của sàn Sendo và đặc sản các miền của sàn Voso. Tổng cộng gồm 180,609 bình luận đến từ cả 3 sàn.

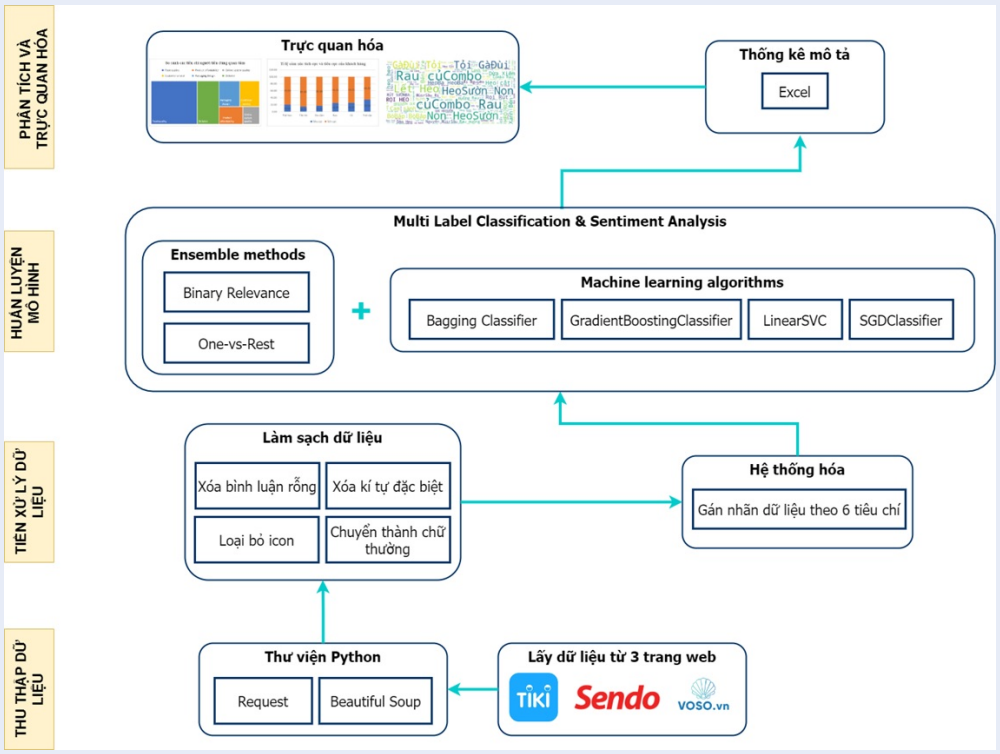
Tiền xử lý dữ liệu

Do tập dữ liệu về bình luận sản phẩm của người dùng cần phải lọc sạch trước khi đưa vào huấn luyện mô hình, để giải quyết vấn đề này chúng tôi sử dụng một số phương pháp xử lý ký tự, từ ngữ trong câu như sửa lỗi chính tả, làm rõ nghĩa của từ. Đồng thời cũng loại bỏ đi icon, ký tự đặc biệt trong câu và xóa tất cả các bình luận rỗng (chỉ gồm các dấu cách khoảng). Cuối cùng là chuyển tất cả câu về chữ thường.

Hệ thống hóa và gán nhãn dữ liệu

Bảng 1 thể hiện các trường dữ liệu được thu thập từ 03 sàn thương mại điện tử. Dữ liệu được hệ thống hóa có các cột với các thông tin bao gồm User_ID (Mã khách hàng bình luận sản phẩm), Item_ID (Mã sản phẩm), Rating (Điểm đánh giá), Comment (Bình luận khách hàng), Product Name (Tên sản phẩm).

Dựa trên công trình của Fang Lyu và Jaewon Choi [6], chúng tôi ghi nhận được nhiều khía cạnh khác nhau về sự quan tâm, yêu cầu và thái độ của khách hàng đối với sản phẩm nông sản và dịch vụ được cung cấp trên các sàn TMĐT, thông qua đó cô đọng thành các nhóm vấn đề mà doanh nghiệp cần quan tâm phát triển hay cải thiện hoạt động kinh doanh của họ trong Bảng 2. Các thẻ chủ đề được nhóm đưa ra trong nghiên cứu là Food quality; Product affordability; Online system quality; Customer service; Packaging design; Delivery. Bên cạnh đó, để có thể bình luận và nghiên cứu sâu hơn, nhóm đã thêm vào yếu tố Tích cực (1) và Tiêu cực (2) cho từng thẻ chủ đề dựa trên nội dung các bình luận của người mua hàng. Với mỗi thẻ chủ đề thì yếu tố tích cực, tiêu cực sẽ có các từ mô tả tương ứng như bảng bên dưới.



Hình 1: Mô hình nghiên cứu tổng quan

Bảng 1: Bảng dữ liệu được hệ thống hóa trước khi gán nhãn

User_ID	Item_ID	Rating	Comment	Product Name
2029473611	48309189	5	Chuyên nghiệp, thân thiện. Đẹp như mô tả. Đóng gói kỹ lưỡng. Chất lượng tuyệt vời. Giá tốt. Giao hàng nhanh bất ngờ. Giao hàng nhanh. Hôm trước đặt hôm sau có luôn. Thời buổi dịch dã giãn cách cứ đặt là nhanh nhất. Bạn giao hàng thân thiện.	[CHỈ GIAO HÀ NỘI] 0.5kg Sả VIETFARM
17579238	107153583	4	Sản phẩm đã nhận được nhanh, đáp ứng nhu cầu cho bữa ăn hàng ngày, cám ơn thịt tươi ngon.	[Chỉ giao HCM] Đùi Tỏi Gà DTP - 1KG
18081630	107153612	3	Thịt hơi mỡ nhưng giao nhanh gọn lẹ. Dịch vầy mà giao hàng khá nhanh.	[Chỉ giao HCM] Ba Chỉ Heo Có Da DTP - 1KG
17642882	114938904	2	Kí khoai được 6 củ thì 3 củ bị sâu và mọc mầm phải bỏ đi.	Khoai lang - 1kg
842046	676057	1	Hàng chưa nhận được mà tin nhắn báo đã giao hàng, bên giao hàng gọi điện kêu tối qua không giao được nên sáng nay giao sớm, sáng giờ cũng chả thấy tin nhắn hay thông báo gì, shop thì nhắn tin mãi không thấy trả lời!!!	Ổi nữ hoàng - 5kg

Bảng 2: Các chủ đề và cụm từ nhận diện

Chủ đề	Mô tả	Tích cực (1)	Tiêu cực (2)
Food quality (FQ)	Chất lượng thực phẩm	Ngon, tươi, ngọt, thơm, chất lượng tốt	Chất lượng cần tốt hơn, hôi, hư, bị dập, héo, chất lượng quá tệ
Product affordability (PA)	Khả năng chi trả sản phẩm/ sản phẩm phù hợp số tiền khách hàng bỏ ra	Giá rẻ, giá tốt	Giá đắt, hơi đắt, giá mắc, hơi mắc
Online system quality (SQ)	Chất lượng của hệ thống trực tuyến/Chất lượng quản lý, vận hành của sàn TMĐT	Hệ thống, luôn tin tưởng [tên sàn TMĐT]	Hệ thống, thất vọng với [tên sàn TMĐT]
Customer service (CS)	Chăm sóc khách hàng/chất lượng phục vụ của bên bán trên sàn TMĐT	Chuyên nghiệp, thân thiện, shop nhiệt tình, lịch sự, trả lời tin nhắn nhanh	Shop cần thân thiện hơn, không phản hồi, thái độ phục vụ chưa tốt
Packaging design (PD)	Đóng gói sản phẩm	Được hút chân không, đóng gói kỹ, đúng trọng lượng	Nên đóng gói kỹ hơn, không có mã vạch, đóng gói sơ xài, hút chân không không được khít/kỹ, thiếu cân
Delivery (DL)	Vận chuyển/giao hàng	Giao hàng nhanh, giao sớm hơn thời gian dự kiến	Giao hàng cần nhanh hơn, giao trễ

Bảng 3: Ví dụ gán nhãn chủ đề dựa theo nội dung đánh giá

Nội dung	Gán nhãn
Tôi chờ cả tuần mà chưa nhận được hàng nhưng trên hệ thống lại thông báo Giao hàng thành công.	SQ2
Thịt tươi, đóng gói cẩn thận. Nhưng kho lên có mùi lạ phải bỏ luôn 1kg thịt.	FQ2PD1
Rau muống rất tươi, rất ngon. Gói hàng rất cẩn thận. Giao hàng rất nhanh và rất có lương tâm trong mùa đại dịch. Rất biết ơn người giao hàng làm việc rất tận tâm. Cầu chúc mọi người nhiều sức khoẻ mùa đại dịch.	FQ1PD1DV1
Sản phẩm không nhãn mác thương hiệu, không hạn sử dụng. Phản ánh thì bảo do đơn hàng quá tải nên nhân viên đóng gói có sai sót, nhưng khi giao bù hàng lần 2 vẫn không có. Giao hàng thì quá hẹn 3-4 ngày mà cứ để thời gian giao hàng trên tiki là trước 16:00 ngày mai. Liên hệ Tiki toàn nói với máy, rồi gặp nhân viên tư vấn thì ngồi đợi dài cổ mới thấy phản hồi.	SQ2CS2PD2DV2
Chuyên nghiệp, thân thiện. Đẹp như mô tả. Đóng gói kỹ lưỡng. Chất lượng tuyệt vời. Giá tốt. Giao hàng nhanh bất ngờ.	FQ1PA1CS1PD1DV1

Nhóm nghiên cứu thực hiện quá trình gán nhãn thủ công dựa trên nội dung bình luận của người mua hàng trên từng sàn TMĐT. Nội dung bình luận sẽ được gán nhãn theo các thẻ chủ đề mà nhóm đã đề xuất bên trên và dựa vào ngữ cảnh của nội dung mà sẽ xét đến các yếu tố tiêu cực và tích cực cho từng trường hợp. Như vậy ứng với từng chủ đề như chất lượng sản phẩm, đóng gói sản phẩm, … sẽ có hai yếu tố tích cực và tiêu cực, tổng cộng là 12 nhãn dán. Tuy nhiên, trong một bình luận, khách hàng có thể đề cập đến nhiều chủ đề và với mỗi chủ đề lại có bình luận tích cực và tiêu cực khác nhau. Do đó, tùy thuộc vào nội dung bình luận, nội dung dán nhãn sẽ khác nhau và không cố định. Bảng 3 trình bày về một số ví dụ chi tiết.

Vector hóa bộ dữ liệu

Để bắt đầu quá trình huấn luyện học máy thì ta không thể đưa dữ liệu đầu vào là văn bản cho máy học mà phải chuyển các dữ liệu này sang dạng vector, một số phương pháp thường được sử dụng vào trong bước này như là: Mô hình Bag of Words, TF-IDF,... Trong bài nghiên cứu này chúng tôi chỉ sử dụng mô hình

TF-IDF.

TF-IDF là kỹ thuật được dùng để tính toán trọng số của các từ. Trọng số này thể hiện tầm quan trọng của một từ trong một văn bản[7].

Mô hình dự đoán

Sau khi thực hiện các mô hình học máy, kết quả dự đoán cần xác định được từng yếu tố tích cực và tiêu cực ứng với mỗi chủ đề. Vì vậy không thể dùng một phương pháp đơn lẻ mà là phải là sự kết hợp của nhiều phương pháp khác nhau. Nhóm nghiên cứu tiến hành xây dựng mô hình kết hợp 2 phương pháp Binary Relevance[8] và One-vs-Rest[9] và các thuật toán học máy như: Bagging Classifier[10], Gradient Boosting Classifier[11], Support Vector Machine (SVM)[12], Stochastic Gradient Descent (SGDClassifier).

Bagging Classifier là thuật toán giúp chúng ta có thể chia tập dữ liệu huấn luyện thành 2 phần ngẫu nhiên và huấn luyện trong mô hình cây quyết định một nửa. Gradient boosting là một phương pháp giúp phát triển mô hình phân lớp và tuyến tính để cải thiện quá trình học của mô hình. Linear SVC là mô hình cho một khả năng xây dựng mô hình đa dạng các tùy biến. SGD ClassifierStochastic Gradient Descent (SGD) là một thuật toán được áp dụng thành công cho các tập dữ liệu quy mô lớn bởi vì việc cập nhật các hệ số được thực hiện cho mỗi trường hợp huấn luyện, thay vì ở cuối các trường hợp. Ngoài ra thì SGD có thể được huấn luyện sử dụng ngắt quãng (sẽ vẫn có thể chạy tiếp được nếu bị tạm dừng).

Mỗi phương pháp sẽ kết hợp với một thuật toán học máy để tiến hành huấn luyện và dự đoán dựa trên tập kiểm thử đã được chia trước đó. Trong mỗi lần vòng lặp chạy sẽ đưa ra thời gian huấn luyện, thời gian dự đoán, điểm số và độ chính xác cho từng mô hình cụ thể.

KẾT QUẢ & THẢO LUẬN

Kết quả được thực nghiệm và đánh giá mô hình được thể hiện trên Bảng 4. Dựa trên các thang đo điểm Precision hay F-score đa số các kết quả cho ra đều trên 80%, đây là một kết quả dự đoán khá cao. Tuy nhiên, khi xét đến khía cạnh thời gian thì thuật toán GradientBoostingClassifier kết hợp với phương pháp Binary Relevance cho ra con số khá cao (4416s cho thời gian huấn luyện) và tương tự với thuật toán BaggingClassifier kết hợp với Binary Relevance (27s cho thời gian dự đoán). Khi xét cùng một tập dữ liệu với nhau mà các thuật toán này cho ra thời gian đã có sự chênh lệch lớn như vậy thì khi tiến hành đưa vào các dữ liệu thực tế có độ lớn gấp nhiều lần đồng nghĩa với việc thời gian cũng sẽ tăng cao.

Thuật toán SGDClassifier với 2 phương pháp kết hợp đều cho kết quả dự đoán ở 2 thang đo Precision và F-score đạt lần lượt là 87% và 82%, tuy nhiên với phương pháp Binary Relevance (BR) thì thời gian huấn luyện cũng như dự đoán lâu hơn (40.5s huấn luyện và 0.5s dự đoán) so với phương pháp One-vs-Rest (OvR) chỉ với 0.2s huấn luyện và 0.003s dự đoán, có thể nói SGDClassifier kết hợp với One-vs-Rest là thuật toán tối ưu nhất cho bài toán và bộ dữ liệu này.

Biểu đồ trong Hình 2 được tạo ra để trực quan hóa thực trạng kinh doanh của các ngành hàng nông sản tính đến tháng 10/2021. Số lượng sản phẩm thu thập được từ các sàn TMĐT là 3032 sản phẩm trải dài trên 6 ngành hàng bao gồm thịt heo, thịt bò, thịt gia cầm, rau, củ, trái cây. Với tổng số khách hàng là hơn 28000 người, và trung bình bình luận của các sàn là xấp xỉ 4.7 sao. Sau khi phân loại các đánh giá, tỷ lệ tích cực ở các lượt bình luận ở khách hàng là rất cao chiếm xấp xỉ 61% trên tổng số.

Số lượng bình luận khách hàng để lại bắt đầu từ các tháng giữa năm của 2021 có những chuyển biến rõ rệt, tăng dần theo thời gian và đạt đỉnh vào tháng 8 cùng năm, tháng 8 là khoảng thời gian dịch bệnh bùng phát mạnh nhất với chiều hướng xấu nhưng lại là cú hích tích cực cho các hoạt động kinh doanh nông sản trên các sàn TMĐT.

Một số tiêu chí như chăm sóc khách hàng, giao hàng, quản lý hệ thống có tỷ lệ tích cực và tiêu cực không quá chênh lệch, điều này cũng có thể lý giải bởi tình hình dịch bệnh làm ảnh hưởng không nhỏ. Khi các lệnh giãn cách được ban hành, đã trở thành rào cản đối với việc đảm bảo thời gian, chất lượng giao hàng của các sàn thương mại điện tử; bên cạnh đó, lượng mua tăng đột biến dẫn đến việc các cửa hàng không bao quát được quá trình chăm sóc khách hàng và hệ thống quá tải phát sinh các trục trặc ảnh hưởng đến trải nghiệm mua hàng của khách hàng.

Có một sự khác biệt lớn khi nhân tố thứ hai mà người tiêu dùng quan tâm là việc vận chuyển sản phẩm, trong khi đối với các nghiên cứu trước đó đã chỉ ra rằng giá cả là ưu tiên hàng đầu khi người tiêu dùng quyết định mua sản phẩm trên các trang TMĐT. Đối với những mặt hàng tươi sống, người tiêu dùng rất quan trọng việc giao hàng đúng giờ, giao chậm thì thịt rau đến tay khách hàng đã mất ngon, thậm chí hư hỏng.

Nhóm nghiên cứu tiến hành thu thập tiếp dữ liệu từ cuối tháng 10/2021 đến tháng 5/2022 và thể hiện qua Hình 3. Tình hình dịch cơ bản đã được kiểm soát trong thời điểm. Mục tiêu của việc nghiên cứu này giúp xác định liệu sau khi dịch bệnh được kiểm soát khách hàng có còn tiếp tục mua hàng nông sản trên

Bảng 4: So sánh các mô hình

Machine learning algorithms	Ensemble methods	Precision	F_score	Thời gian huấn luyện	Thời gian dự đoán
BaggingClassifier	BR	84%	80%	1808.286	26.931
	OvR	84%	80%	123.602	0.222
GradientBoostingClassifier	BR	87%	79%	4416.853	1.361
	OvR	87%	79%	72.042	0.035
LinearSVC	BR	86%	82%	8.118	0.581
	OvR	86%	82%	0.458	0.004
SGDClassifier	BR	87%	82%	40.513	0.558
	OvR	87%	82%	0.221	0.003

Hình 2: Thực trạng kinh doanh đến tháng 10/2021 của các ngành hàng nông sản

các sàn TMĐT hay không và phản hồi của họ như thế nào.

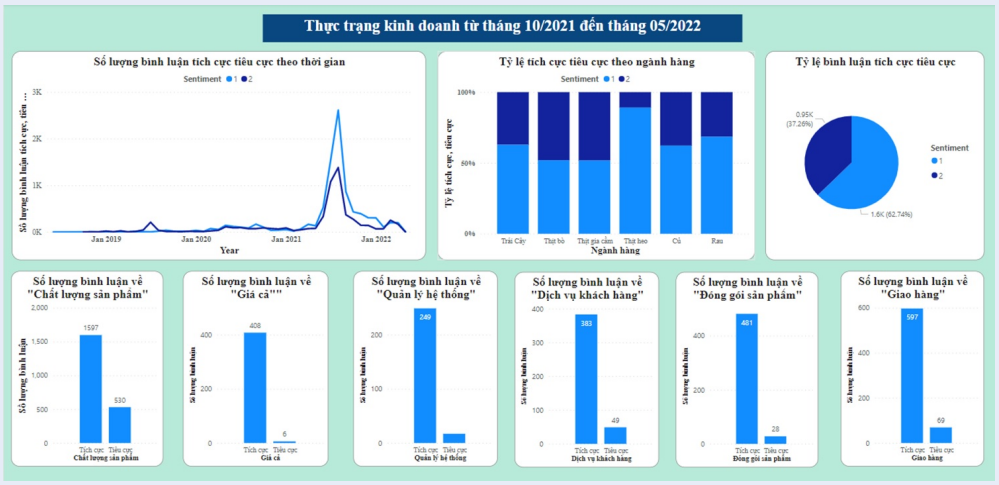
Tổng số bình luận thu về được là 2547 bình luận, sau khi phân loại các đánh giá, tỷ lệ tiêu cực ở các lượt bình luận ở khách hàng là rất cao trên 62% trên tổng số. Từ tháng 2/2022, số lượng bình luận tích cực và tiêu cực xấp xỉ nhau, cho thấy sau dịch tình hình nông sản trên các sàn TMĐT có chiều hướng đi xuống. Ngược lại với những bình luận trước tháng 11/2021, thịt bò và thịt gia cầm là hai ngành hàng có số bình luận tiêu cực cao nhất với 48%, trước đó thịt bò chỉ có 134% bình luận tiêu cực. Thịt heo là ngành có số lượng bình luận tiêu cực thấp nhất, xấp xỉ 11%.

Trong nghiên cứu tiếp theo chúng tôi sẽ tiến hành cài đặt hệ thống để tự động cập nhật dữ liệu. Dữ liệu sẽ tự trích xuất dữ liệu từ trên website các sàn TMĐT, loại bỏ dữ liệu trùng lặp trước khi lưu vào cơ sở dữ liệu và tự động đưa vào biểu đồ để trực quan hóa.

KẾT LUẬN

Bài nghiên cứu đã giải quyết được ba vấn đề quan trọng đóng góp về mặt khoa học và thực tiễn trong lĩnh vực phân tích dữ liệu bình luận khách hàng:

Thứ nhất, nghiên cứu đã tìm ra những khía cạnh nào ảnh hưởng đến trải nghiệm người dùng về nông sản trên các sàn TMĐT. Mức độ mà những khía cạnh này



Hình 3: Thực trạng kinh doanh từ tháng 10/2021 đến tháng 05/2022 của các ngành hàng nông sản

ảnh hưởng đến cảm xúc của người mua được coi là tích cực hay tiêu cực. Các khía cạnh bao gồm: chất lượng sản phẩm, đóng gói sản phẩm, vận chuyển, giá thành, dịch vụ khách hàng và hệ thống đặt hàng.

Thứ hai, nghiên cứu đã phát triển mô hình học máy tự động phát hiện và gán nhãn cho tất cả các chủ đề cùng xuất hiện trong một câu đánh giá. Mô hình cũng tự động gán nhãn cảm xúc cho các bình luận.

Thứ ba, xây dựng các biểu đồ và thể hiện trực quan số liệu phân tích, giúp doanh nghiệp có chiến lược phát triển các sản phẩm từ nông sản và tham gia thị trường tốt hơn. Dựa vào đó mà doanh nghiệp có thể phân tích điểm mạnh và điểm yếu của mình cũng như đối thủ cạnh tranh để có các chiến lược tốt hơn.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Kinh tế - Luật, ĐHQG-HCM và nhóm tác giả gửi lời cảm ơn đến anh Võ Trần Đông Dương, sinh viên khoa Khoa học và Kỹ thuật thông tin, trường Đại học Công Nghệ Thông Tin, ĐHQG-HCM đã hỗ trợ nhóm.

DANH MỤC TỪ VIẾT TẮT

TMĐT - Thương mại điện tử;

NLP: Natural Language Process (Xử lý ngôn ngữ tự nhiên);

TF-IDF: Term Frequency–Inverse Document Frequency;

HTTP: Hypertext Transfer Protocol (Giao thức truyền siêu văn bản);

JSON: JavaScript Object Notation (Một kiểu dữ liệu mở trong Javascript);

API: Application Programming Interface (Giao diện lập trình ứng dụng);

BR: Binary Relevance;

OvR: One-vs-Rest.

XUNG ĐỘT LỢI ÍCH

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

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Các tác giả cùng đóng góp về ý tưởng, mục tiêu, lựa chọn phương pháp nghiên cứu, thảo luận các kết quả nghiên cứu và các vấn đề liên quan đến trực quan hoá dữ liệu và kết quản nghiên cứu.

TÀI LIỆU THAM KHẢO

1. Humphreys A, Wang RJ. Automated text analysis for consumer research. Journal of Consumer Research. 2018 Apr 1;44(6):1274-306;Available from: https://doi.org/10.1093/jcr/ucx104.
2. Tao D, Yang P, Feng H. Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive reviews in food science and food safety. 2020 Mar;19(2):875-94;PMID: 33325182. Available from: https://doi.org/10.1111/1541-4337.12540.
3. Balazs JA, Velásquez JD. Opinion mining and information fusion: a survey. Information Fusion. 2016 Jan 1;27:95-110;Available from: https://doi.org/10.1016/j.inffus.2015.06.002.
4. Hassan A, Abbasi A, Zeng D. Twitter sentiment analysis: A bootstrap ensemble framework. In2013 international conference on social computing 2013 Sep 8 (pp. 357-364). IEEE;PMID: 24734318. Available from: https://doi.org/10.1109/SocialCom.2013.56.
5. Wang G, Sun J, Ma J, Xu K, Gu J. Sentiment classification: The contribution of ensemble learning. Decision support systems. 2014 Jan 1;57:77-93;Available from: https://doi.org/10.1016/j.dss.2013.08.002.
6. Lyu F, Choi J. The forecasting sales volume and satisfaction of organic products through text mining on web customer reviews. Sustainability. 2020 Jan;12(11):4383;Available from: https://doi.org/10.3390/su12114383.

7. Yun-tao Z, Ling G, Yong-cheng W. An improved TF-IDF approach for text classification. Journal of Zhejiang University-Science A. 2005 Aug;6(1):49-55;Available from: https://doi.org/10.1631/jzus.2005.A49.
8. Zhang ML, Li YK, Liu XY, Geng X. Binary relevance for multi-label learning: an overview. Frontiers of Computer Science. 2018 Apr;12(2):191-202;Available from: https://doi.org/10.1007/s11704-017-7031-7.
9. Ramírez J, Górriz JM, Ortiz A, Martínez-Murcia FJ, Segovia F, Salas-Gonzalez D, Castillo-Barnes D, Illán IA, Puntonet CG, Alzheimer's Disease Neuroimaging Initiative. Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. Journal of neuroscience methods. 2018 May 15;302:47-57;PMID: 29242123. Available from: https://doi.org/10.1016/j.jneumeth.2017.12.005.
10. Dong L, Yuan Y, Cai Y. Using Bagging classifier to predict protein domain structural class. Journal of biomolecular structure & dynamics. 2006 Dec 1;24(3):239-42;.
11. Lusa L. Gradient boosting for high-dimensional prediction of rare events. Computational Statistics & Data Analysis. 2017 Sep 1;113:19-37;Available from: https://doi.org/10.1016/j.csda.2016.07.016.
12. Noble WS. What is a support vector machine?. Nature biotechnology. 2006 Dec;24(12):1565-7;PMID: 17160063. Available from: https://doi.org/10.1038/nbt1206-1565.

Science & Technology Development Journal – Economics - Law and Management 2022, 6(4):3682-3690

Open Access Full Text Article

Research article

Applying machine learning methods to analyze customer comments about fresh food on e-commerce platforms in Vietnam

Anh Nguyen Thi Ngoc, Giang Phan Thi Ha, Giang Vo Chi, An Nguyen Ba Thinh, Dat Nguyen Phat, Ai Ho Thy Nhan, Hung Nguyen Quang*

ABSTRACT

In recent years, farmers have developed the sale of agricultural products toward consumers via e-commerce platforms. E-commerce has become a new and effective way to help farmers access the market. Thus, in comparison to other commodities, agricultural products are heavily affected by seasonality, with complex factors such as short shelf life, vulnerability to damage, and high transportation costs. Consumers set high standards for the quality, speed of delivery, frequency of consumption, and unit price of these products. Analyzing customer reviews helps businesses discover consumer decision-making mechanisms, thereby forming an appropriate marketing strategy for their agricultural products. Besides, they will see what customers are unsatisfied with to solve and improve the quality of products and services. In this study, the authors research and propose machine research methods to classify and screen customers' comments about agricultural products on three e-commerce platforms: Tiki, Sendo and Voso. Experimenting with the model on the collected data set with the results of the sgdclassifier algorithm combined with the One-vs-Rest method gave the best prediction results with 87%. The study also builds charts and directly shows the amount of data analyzing the factors affecting customer satisfaction with quality products as well as seller's services and e-commerce platforms. In addition, the study proposes recommendations to help businesses improve the quality of products and services, thereby providing better strategies to attract and retain customers.

Key words: user comments, machine learning, agricultural products, sentiment analysis, e-commerce

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

Correspondence

Hung Nguyen Quang, University of Economics and Law, Ho Chi Minh City, Vietnam Vietnam National University, Ho Chi Minh City, Vietnam

Email: hungnq@uel.edu.vn

History

- Received: 18-9-2022
- Accepted: 16-12-2022
- Published: 31-1-2023

DOI : https://doi.org/10.32508/stdjelm.v6i4.1132

Cite this article : Ngoc A N T, Ha G P T, Chi G V, Thinh A N B, Phat D N, Nhan A H T, Quang H N. **Applying machine learning methods to analyze customer comments about fresh food on e-commerce platforms in Vietnam**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 2022, 6(4):3682-3690.