PHƯƠNG PHÁP PHÂN LOẠI DỮ LIỆU BÌNH LUẬN CỦA KHÁCH HÀNG TRỰC TUYẾN VIỆT NAM DỰA VÀO HỌC MÁY CÓ GIÁM SÁT

METHODS FOR CLASSIFYING COMMENT DATA OF ONLINE CUSTOMER IN VIETNAM BASED ON SUPERVISED MACHINE LEARNING

Lê Triệu Tuấn[1,*], Đàm Thị Phương Thảo[1]

TÓM TẮT

Nghiên cứu nhằm mục đích ứng dụng phương pháp học máy có giám sát vào việc phân loại dữ liệu là các nội dung bình luận sản phẩm của khách hàng trong mua sắm trực tuyến. Nghiên cứu tiến hành thu thập dữ liệu tự động với 2530 nội dung bình luận của khách hàng về các sản phẩm trên các trang thương mại điện tử hàng đầu tại Việt Nam, sau đó thực hiện huấn luyện với các mô hình học máy có giám sát để tìm ra mô hình phù hợp nhất với bộ dữ liệu huấn luyện và áp dụng mô hình này để dự báo nội dung nhận xét cho toàn bộ tập dữ liệu. Kết quả cho thấy các phương pháp học máy Support Vector Machines (SVM), Decision Tree (DT) và Neural Network (NN) có hiệu suất tốt nhất với việc phân loại nhận xét của khách hàng bằng Tiếng Việt. Kết quả nghiên cứu có giá trị tham khảo cho các ứng dụng khai thác nội dung nhận xét trong lĩnh vực kinh doanh trực tuyến.

Từ khóa: *Khai thác dữ liệu bình luận, phân loại bình luận, phân loại bằng học máy có giám sát, khai phá dữ liệu, dữ liệu lớn.*

ABSTRACT

The study aims to apply a supervised machine learning method to classify comments data as customer product comments in online shopping. The study conducted automatic data collection with 2,530 customer comments about products on the top of e-commerce sites in Vietnam, then trained with supervised machine learning models. to find the model that the best fits the training dataset and apply this model to predict the comment content for the entire dataset. The results show that the Machine Learning methods Support Vector Machines (SVM), Decision Tree (DT) and Neural Network (NN) have the best performance with classifying customer comments in Vietnamese. The research results have reference value for comment mining applications in the field of online business.

Keywords: *Comments mining, comment data classification, classification by supervised machine learning, data mining, big data.*

[1]Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên
Email: lttuan@ictu.edu.vn
Ngày nhận bài: 25/11/2021
Ngày nhận bài sửa sau phản biện: 05/01/2022
Ngày chấp nhận đăng: 25/02/2022

1. GIỚI THIỆU

Hàng ngày, trên các trang thương mại điện tử có rất nhiều những nội dung bình luận, nhận xét của khách hàng về các sản phẩm hoặc dịch vụ. Việc phân tích thống kê lại xem những nội dung bình luận, nhận xét đó là tích cực hay tiêu cực sẽ giúp cho doanh nghiệp biết được chất lượng sản phẩm, chất lượng phục vụ, tâm lý khách hàng và từ đó đưa ra những thay đổi trong kinh doanh [6].

Với sự bùng nổ của dữ liệu lớn (Big Data) hiện nay, việc khai thác các nội dung bình luận, nhận xét của khách hàng theo cách truyền thống là điều không thể. Mà các dữ liệu này cần được thu thập và khai thác tự động, cho phép các nhà kinh doanh theo dõi hành vi mua sắm, phát hiện sở thích và hỗ trợ khách hàng mua các sản phẩm, dịch vụ một cách tốt nhất [5].

Phân loại nội dung là một bước quan trọng trong phương pháp học máy (Machine Learning) để nghiên cứu và khai thác nội dung bình luận, nhận xét của khách hàng trực tuyến. Đã có nhiều công trình nghiên cứu về phương pháp phân loại nội dung ở nhiều mức độ khác nhau, và từ kết quả tìm hiểu từ các công trình nghiên cứu trong và ngoài nước, tác giả nhận thấy có hai cách tiếp cận trong phân loại nội dung bình luận, nhận xét trực tuyến theo phương pháp học máy: (1) Học máy có giám sát (Supervised Machine Learning) và (2) Học máy không giám sát (Unsupervised Machine Learning). Nghiên cứu về phương pháp khai thác nội dung bình luận, nhận xét của khách hàng trực tuyến không phải mới. Tuy nhiên, mỗi phương pháp có những ưu và nhược điểm riêng, không có phương pháp nào được xem là chính xác tuyệt đối. Nghiên cứu này áp dụng phương pháp học máy có giám sát để thực hiện phân loại các nội dung bình luận, nhận xét trực tuyến với nguồn dữ liệu được thu thập tự động, trong đó với 2530 các bình luận, nhận xét của khách hàng về các sản phẩm trên các trang thương mại điện tử [3].

2. CƠ SỞ LÝ THUYẾT

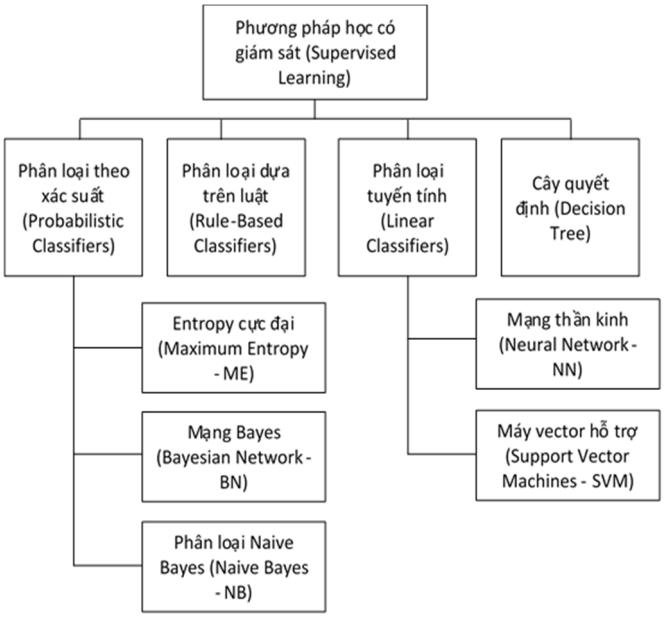
2.1. Khai thác bình luận của khách hàng

Khai thác bình luận của khách hàng là lĩnh vực nghiên cứu nhằm phân tích, đánh giá nhận định của khách hàng về các đối tượng như: Sản phẩm, dịch vụ, tổ chức, cá nhân, sự kiện, chủ đề và các thuộc tính của chúng [4, 8]. Một quy trình khai thác thường gồm ba bước chính: (1) Thu thập dữ

liệu (Comment Retrieval), (2) Phân loại nội dung bình luận (Comment Classification), và (3) Tổng hợp nhận xét (Comment Summarization) [1, 2]. Trong đó, phân loại được coi là bước quan trọng nhất nhằm mục đích phân lớp bình luận theo các mức: Lạc quan (Positive); tiêu cực (Negative). Theo [4], khai thác bình luận được chia thành ba mức độ: (1) Mức tài liệu (Document Level), ở mức khai thác này, giả định mỗi tài liệu thể hiện nội dung bình luận về một thực thể đơn. Vì vậy, các phân tích sẽ không thể áp dụng được cho những tài liệu đề cập đến nhiều đối tượng; (2) mức câu (Sentence Level), ở mức khai thác này, giả định mỗi câu thể hiện nội dung về một đối tượng, tuy nhiên, các phân tích sẽ bỏ qua những câu có nhiều mệnh đề, mỗi mệnh đề thể hiện nhận xét về các đối tượng khác nhau; và (3) mức thực thể, khía cạnh (Entity/Aspect Level), thay vì khai thác nhận xét theo cấu trúc ngôn ngữ (tài liệu, câu, mệnh đề...), mức phân tích này xem xét nội dung theo mục tiêu (Target), mục tiêu của bình luận có thể là đối tượng hoặc khía cạnh (thuộc tính) của đối tượng. Ngày nay, với sự bùng nổ của dữ liệu lớn, việc khai thác các bình luận của khách hàng trở thành mối quan tâm lớn của các nhà kinh doanh, đặc biệt là các công ty có website cho phép người dùng được bình luận, nhận xét trên đó. Khai thác bình luận cũng có thể được bổ sung cho các hệ thống tư vấn mua hàng (Recommender Systems) để đề xuất các sản phẩm được nhận xét tích cực và không nên giới thiệu các danh mục sản phẩm nhận được nhiều nhận xét tiêu cực [7, 8].

2.2. Phân loại bình luận bằng phương pháp máy học có giám sát

Phương pháp học có giám sát là một kỹ thuật của ngành Khoa học máy tính để xây dựng một hàm từ dữ liệu huấn luyện. Dữ liệu huấn luyện bao gồm các cặp gồm đối tượng đầu vào (thường dạng vec-tơ), và đầu ra mong muốn. Đầu ra của một hàm là dự đoán một nhãn cho một đối tượng [9].



Hình 1. Các phương pháp trong học máy có giám sát

3. PHƯƠNG PHÁP NGHIÊN CỨU

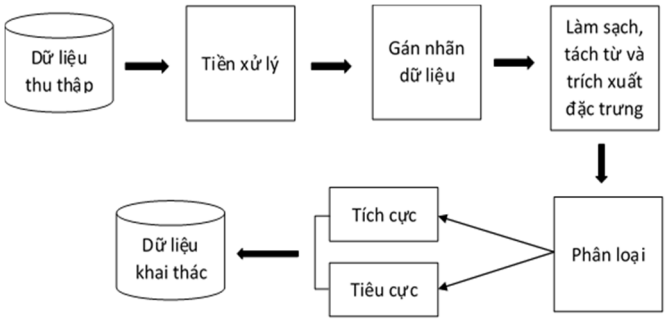
Nghiên cứu này được tiến hành theo phương pháp khai phá tri thức từ dữ liệu KDD (Knowledge Discovery in Databases). Các bước trong quy trình nghiên cứu được thực hiện như trong hình 2. Môi trường thực nghiệm được cài đặt bằng ngôn ngữ lập trình Python với sự hỗ trợ của công cụ tách từ Python Vietnamese Toolkit (dành cho ngôn ngữ tiếng Việt) và các thư viện có sẵn.

Bước 1. Thu thập và tiền xử lý dữ liệu

Nghiên cứu đã tiến hành thu thập dữ liệu bằng chương trình tự động, dữ liệu lấy từ các trang thương mại điện tử, như: Lazada.vn; Tiki.vn. Đây là phương pháp thu thập nội dung tự động từ các trang HTML của bất kỳ tài nguyên Internet bằng các chương trình hoặc mã lệnh đặc biệt. Với đối tượng và phạm vi nghiên cứu hướng đến là ngôn ngữ tiếng việt. Do đó, dữ liệu chỉ sử dụng những bình luận của khách hàng bằng tiếng Việt. Tiếp đến, nghiên cứu đã tiến hành tiền xử lý dữ liệu bằng cách loại bỏ những dữ liệu khuyết, những nhận xét không chứa đựng thông tin cần thiết để tiến hành bước xử lý tiếp theo [1].

Bước 2. Gán nhãn dữ liệu (Data Labeling)

Bước 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. Thông thường đối với các nghiên cứu ứng dụng phương pháp máy học, tập dữ liệu này sẽ được xây dựng bằng thủ công. Tuy nhiên, trong nghiên cứu này, sau khi xem xét ngẫu nhiên nội dung của tập dữ liệu bình luận, nhận xét đã thu thập được và dựa vào kết quả điểm đánh giá (trường rating trong tập dữ liệu), nghiên cứu này nhận thấy các phản hồi có điểm đánh giá nhỏ hơn 7,0 mang ý nghĩa tiêu cực (Negative), và ngược lại, các phản hồi có điểm đánh giá lớn hơn 7,0 mang ý nghĩa tích cực (Positive). Do đó, tập dữ liệu huấn luyện được xác định có 2530 phản hồi, trong đó có 89 bình luận là tiêu cực (được gán nhãn 0) và 2441 bình luận là tích cực (được gán nhãn 1).



Hình 2. Quy trình nghiên cứu

Bước 3. Làm sạch, tách từ và trích xuất đặc trưng

Làm sạch dữ liệu (Data Cleaning): Bước này tiến hành làm sạch dữ liệu trước khi bắt đầu xử lý trên tập dữ liệu, bao gồm một số công đoạn xử lý ngôn ngữ tự nhiên như loại bỏ hư từ (Stop Words), hoặc kiểm tra chính tả…

Tách từ (Words Segmentation): Bước này rất quan trọng trong việc xử lý ngôn ngữ tự nhiên, và đặc biệt đối với ngôn ngữ Tiếng Việt vì có nhiều từ ghép, tách từ theo nhiều cách

khác nhau có thể sẽ gây ra sự nhập nhằng về mặt ngữ nghĩa. Nghiên cứu này kế thừa bộ thư viện tách từ Python Vietnamese Toolkit.

Trích xuất đặc trưng (Feature Extraction): Bước này sẽ 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 (Input) cho thuật toán phân loại. Nghiên cứu này lựa chọn từ khóa theo phương pháp TF-IDF (Term Frequency/Inverse Document Frequency), giá trị TF-IDF của một từ khóa là một con số thu được qua thống kê thể hiện mức độ quan trọng của từ khóa này trong một bình luận. TF-IDF của từ khóa w_i trong phản hồi *d* được tính bằng công thức sau:

$$if\_idf_{id} = f_{id} \times \log \frac{N}{n_i}$$

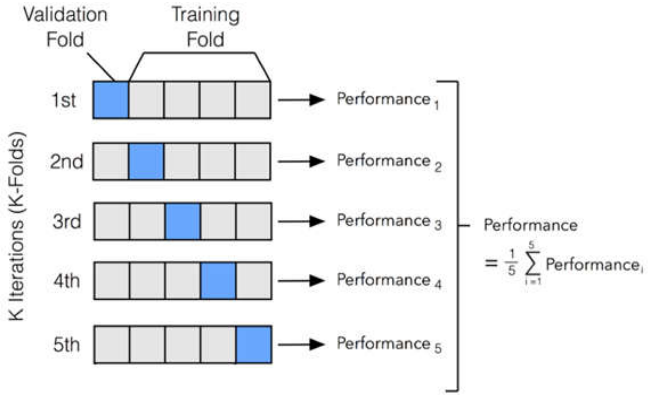
Trong đó

f_{id}: Tần suất xuất hiện của từ khóa w_i trong nhận xét d

N: Tổng số nhận xét

n_i: Số nhận xét mà có từ khóa w_i xuất hiện

Huấn luyện mô hình phân loại bình luận: Giai đoạn này nhằm mục đích xác định một bình luận, nhận xét của khách hàng là "tích cực" hay "tiêu cực". Nghiên cứu này ứng dụng các thuật toán phân loại "Máy vector hỗ trợ - SVM" thuộc nhóm máy học giám sát (Supervised Machine Learning) được cho là tốt nhất, thuật toán Naive Bayes, Neural Network và Decision Tree. Dựa trên kết quả tổng hợp từ các nghiên cứu trước có liên quan đến đề tài để tìm ra mô hình phù hợp nhất đối với tập dữ liệu là các nhận xét đã được phân loại. Từ đó, tiến hành dự báo cho các dữ liệu nhận xét chưa được phân loại hoặc các dữ liệu nhận xét mới phát sinh mà không cần phải huấn luyện lại. Quá trình huấn luyện được tiến hành bởi phương pháp kiểm tra chéo k-fold (k-fold cross validation), chia ngẫu nhiên dữ liệu thành K tập con không giao nhau [10]. Mỗi tập thực nghiệm (trong số K lần), một tập con được sử dụng làm tập kiểm thử, và (K-1) tập con còn lại được dùng làm tập huấn luyện. Nghiên cứu này sử dụng K = 5.



(Nguồn: internet)

Hình 3. Phương pháp K-Fold (Sarvesh Harikant)

Trong đó:

Performance: Hiệu suất trung bình của 5 lần thực nghiệm

K Iterations: Lặp lại K lần

Validation Fold: Tập dữ liệu dùng để kiểm thử

Training Fold: Tập dữ liệu dùng để huấn luyện

Bước 4. Đánh giá hiệu quả phân loại

Nghiên cứu sử dụng phương pháp đánh giá mô hình phân loại là dựa trên các chỉ số tính toán trong ma trận nhầm lẫn (Confusion Matrix) như bảng 1.

Bảng 1. Ma trận nhầm lẫn (Confusion Matrix)

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

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

Trong đó:

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision \times Recall}$$

4. KẾT QUẢ NGHIÊN CỨU

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

Kết quả thu thập dữ liệu được 2530 bình luận, nhận xét sản phẩm bằng tiếng Việt của 15 mặt hàng trên 5 website thương mại điện tử khác nhau tại Việt Nam. Dữ liệu được phân bố như trong bảng 2.

Bảng 2. Kết quả thu thập và tiền xử lý dữ liệu

STT	Mặt hàng	Số lượng	Số lượt phản hồi	Trung bình
1	Tivi	14	173	12,4
2	Tủ lạnh	6	93	15,5
3	Điều hòa	9	246	21,8
4	Quần Jean	3	12	4,0
5	Áo thun	4	104	13,5
6	Quần bơi nam	3	45	15,0
7	Điện thoại Iphone 12	2	335	123,0
8	Điện thoại Iphone 11 pro max	2	248	99,0
9	Điện thoại Iphone 10	4	74	18,5
10	Điện thoại Samsung Galaxy A32	3	297	99,0
11	Điện thoại Samsung A52	3	207	52,3
12	Điện thoại Samsung A72	2	65	32,5
13	Điện thoại OPPO Watch	5	289	57,8
14	Điện thoại OPPO Reno5	4	32	8,0
15	Điện thoại OPPO A53	4	310	77,5
Tổng		**68**	**2530**	

4.2. Kết quả huấn luyện và đánh giá mô hình phân loại

Nghiên cứu tiến hành huấn luyện bằng 04 thuật toán, bao gồm: Naive Bayes (NB), Support Vector Machines (SVM), Neural Network (NN), Decision Tree (DT). Hiệu quả huấn luyện của các thuật toán được thể hiện trong bảng 3.

Bảng 3. Độ chính xác của các thuật toán được huấn luyện theo phương pháp K-Fold (K = 5)

STT	Tên phương pháp	Độ chính xác trung bình	Độ lệch chuẩn	Thời gian huấn luyện (giây)
1	Naive Bayes (NB)	0,48	0,05	16,02
2	Support Vector Machines (SVM)	0,80	0,02	4,33
3	Neural Network (NN)	0,76	0,03	312,29
4	Decision Tree (DT)	0,70	0,03	315,56

Bảng kết quả cho thấy mô hình SVM có độ chính xác khá cao (0,80), mô hình NN (0,76) và mô hình DT (0,70). Nghĩa là các mô hình này tương đối phù hợp với tập dữ liệu huấn luyện. Và đồng thời, xét thêm yếu tố thời gian huấn huyện thì mô hình SVM có thời gian huấn luyện thấp nhất. Do đó, các ứng dụng tiếp theo có thể dùng mô hình này như một công cụ để phân loại nội dung cho các dữ liệu bình luận, nhận xét chưa được phân loại hoặc các dữ liệu bình luận mới phát sinh mà không cần phải huấn luyện lại. Kết quả nghiên cứu này đã giúp xác định phương pháp và công cụ phân loại nội dung bình luận phù hợp.

5. KẾT LUẬN

Nghiên cứu này đã tiến hành lược khảo cơ sở lý thuyết về phương pháp phân loại bình luận và đề xuất ứng dụng phương pháp máy học có giám sát cho việc khai thác nội dung bình luận một cách tự động. Kết quả thực nghiệm cho thấy phương pháp Support Vector Machines (SVM) là tốt nhất trong các phương pháp huấn luyện. Nghiên cứu này có giá trị tham khảo cho các ứng dụng khai thác nội dung bình luận trong lĩnh vực bán hàng trực tuyến. Tuy nhiên, nghiên cứu này vẫn còn nhiều hạn chế, có thể tiếp tục thực hiện trong thời gian tới hoặc trong những nghiên cứu tiếp theo: Thứ nhất, về thu thập dữ liệu, nghiên cứu này thu thập dữ liệu là các bình luận, nhận xét của khách hàng về các mặt hàng trên một số trang thương mại điện tử chứ chưa thu thập được trên hầu hết các trang; Thứ hai, về thang đo, nghiên cứu này chỉ phân loại phản hồi khách hàng theo thang đo 2 mức: Tích cực (Positive), và tiêu cực (Negative). Hướng nghiên cứu kế tiếp có thể sử dụng thang đo nhiều mức hơn (ví dụ theo thang đo Likert 5 mức); Thứ ba, về kỹ thuật phân loại nội dung nhận xét, nghiên cứu này chỉ sử dụng phương pháp máy học có giám sát, nếu kết hợp với phương pháp từ vựng dựa trên ngữ nghĩa có thể sẽ cho kết quả tốt hơn.

TÀI LIỆU THAM KHẢO

[1]. K.M. Kavitha, et al., 2020. *Analysis and Classification of User Comments on YouTube Videos.* International Workshop on Artificial Intelligence for Natural Language Processing (IA&NLP 2020), Vol 177, pp. 593-598.

[2]. Kumar S., Reddy B., 2016. *An analysis on opinion mining: Techniques and tools*. Indian Journal of Research, 5(8), pp. 489–492.

[3]. Le N. M., Do B. N., Nguyen V. D., Nguyen T. D., 2013. *VNLP: An open source framework for Vietnamese natural language processing.* In Proceedings of the Fourth Symposium on Information and Communication Technology, 88–93.

[4]. Liu B., 2012. *Sentiment analysis and opinion mining.* Synthesis Lectures on Human Language Technologies, 5(1), pp. 1–167.

[5]. Mehdi Golzadeh, et al., 2021. *A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments.* Journal of Systems and Software, Vol 175 pp. 110-125.

[6]. Ochilbek Rakhmanov, 2020. *A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments.* Procedia Computer Science, Vol 178, pp. 194-204.

[7]. Özlem, Tutku, 2021. *Classification of rare diseases; A comment on 'atlas of esophageal atresia'*. Journal of Pediatric Surgery.

[8]. Pang B., Lee L., 2008. *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval, 2(1-2), pp. 1–135.

[9]. Reynaldo, et al., 2019. *Gender Demography Classification on Instagram based on User's Comments Section*. 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 157, 64-71.

[10]. Sarvesh Harikant, 2021. *K Fold Cross Validation Technique*. Retrieved 6/2021 from https://inblog.in/K-Fold-Cross-Validation-Technique-NCaSQ8Kmfh.

AUTHORS INFORMATION

Le Trieu Tuan, Dam Thi Phuong Thao

Thainguyen University of Information and Communication Technology