Nhận Diện Cảm Xúc và Phân Loại Chủ Đề Thông Qua Bình Luận Của Khách Hàng

1st Võ Hoàng An
Trường đại học công nghệ thông tin
21520555@gm.uit.edu.vn

2nd Mai Khánh Linh
Trường đại học công nghệ thông tin
21522287@gm.uit.edu.vn

3rd Trần Thái Hoà
Trường đại học công nghệ thông tin
21522082@gm.uit.edu.vn

Tóm tắt nội dung*—Trong những năm gần đây, dưới sự phát triển nhanh chóng của công nghệ 4.0 và tác động mạnh mẽ của đại dịch Covid-19, kinh doanh trực tuyến ngày càng được mở rộng, góp phần hình thành và thúc đẩy thói quen mua sắm thông qua các sàn thương mại điện tử. Do đó, việc hiểu rõ hành vi của khách hàng thông qua cảm xúc của khách hàng về sản phẩm và dịch vụ trở thành một vấn đề quan trọng. Bài toán nhận diện cảm xúc và mô hình chủ đề của người tiêu dùng thông qua các bình luận đang được nhiều doanh nghiệp quan tâm và tìm kiếm các phải pháp. Vì thế việc xây dựng các mô hình học máy, các phương pháp tiền xử lý nhằm phát triển các chiến lược thích hợp để thỏa mãn trải nghiệm mua sắm của khách hàng vô cùng quan trọng đối với các doanh nghiệp. Vì những lý do đó, chúng tôi xây dựng bộ dữ liệu STiCS do chúng tôi tự thu thập từ sàn thương mại điện tử Shopee để thực nghiệm các mô hình học máy với mục tiêu tìm ra mô hình tốt nhất cho việc nhận diện cảm xúc và tìm ra các chủ đề tiềm ẩn dựa trên bình luận của khách hàng. Sau khi thực nghiệm chúng tôi thấy rằng đối với bài toán phân loại cảm xúc, mô hình kết hợp từ ba mô hình SVC, Logistic regression và LinearSVC cho kết quả tốt nhất với 80,84% với độ đo accuracy và 71,60% với độ đo $F1_{macro}$, còn đối với bài toán mô hình chủ đề, mô hình NMF với số lượng chủ đề là năm cho ra kết quả tốt nhất.

***Từ khoá*—sentiment analysis, topic modeling, machine learning, classification, LDA, TF-IDF vectorizer, count vectorizer**

I. Giới thiệu

Sau đại dịch Covid 19, thương mại điện tử phát triển bùng nổ và trở thành đầu tàu của kinh tế số, góp phần thúc đẩy thói quen tìm kiếm sản phẩm trực tiếp thông qua các ứng dụng mua sắm trực tuyến. Theo trang tin datareportal[1], khi tiến hành khảo sát người dụng Internet trong độ tuổi 16-64, có tới 58,4% người dùng Internet có mua hàng hóa hoặc dịch vụ; 28,3% người mua hàng thông qua cửa hàng trực tuyến; 24,6% người sử dụng dịch vụ so sánh giá trực truyến. Ngoài ra, theo "Sách trắng thương mại điện tử Việt Nam năm 2022", tỷ lệ người dùng Internet tham gia mua sắm trực tuyến tăng mạnh, lên đến khoảng 74,8%. Thông qua các con số được thống kê và các dự đoán về sự phát triển của thương mại điện tử trong những năm tiếp theo, có thể thấy được việc nắm bắt thị hiếu của người tiêu dùng online chiếm vai trò quan trọng trong việc nâng cao chất lượng về sản phẩm-dịch vụ, cũng như hỗ trợ các doanh nghiệp hình thành thương hiệu, tạo dựng niềm tin và thu hút sự chú ý của khách hàng.

Bên cạnh các bình luận có đầy đủ thông tin, mang lại nhiều giá trị thì hầu hết các đánh giá trên các sàn thương mại điện tử đều là dạng dữ liệu không có cấu trúc, không được chuẩn hóa, chứa nhiều từ vô nghĩa gây ảnh hưởng tiêu cực đến bộ dữ liệu và quá trình thực hiện các mô hình học máy. Do đó, để tăng chất lượng của dữ liệu và hiệu quả của mô hình học máy, việc xử lý dữ liệu chiếm một vai trò vô cùng quan trọng và cần thiết.

Trong bài báo cáo này, chúng tôi kết hợp giữa phân tích cảm xúc (Sentiment Analysis) và mô hình chủ đề (Topic Modeling) để xây dựng mô hình học máy. Bước đầu tiên, chúng tôi tiến hành thu thập các bình luận của người tiêu dùng trên website Shopee.vn[2], một trang thương mại điện tử hàng đầu trong lĩnh vực đặt hàng trực tuyến. Tiếp theo, chúng tôi áp dụng các phương pháp tiền xử lý dữ liệu. Cuối cùng, dựa trên sắc thái của câu bình luận mà phân loại nhãn cảm xúc thành 3 nhãn gồm: tích cực (positive), tiêu cực (negative), trung lập (neutral), đồng thời tìm ra các chủ đề tiềm ẩn trong câu bình luận đó. Ba đóng góp chính của bài báo cáo nào được đúc kết như sau:

1) Chúng tôi giới thiệu bộ dữ liệu Sentiment analysis - Topic modeling in Comment Shopee (STiCS) bao gồm 7224 bình luận và đánh giá được thu thập từ Shopee.
2) Trình bày các phương pháp tiếp cận học máy để phân loại cảm xúc và các phương pháp học giám sát để phân loại chủ đề của bình luận.
3) Phân tích các trường hợp sai, các hạn chế, và các hướng có thể phát triển trong tương lai của bài toán.

Phần còn lại của bài báo cáo được sắp xếp như sau. Đầu tiên, trong phần 2, chúng tôi tiến hành định nghĩa bài toán, mô tả khái quát các nghiên cứu đã được thực hiện về công trình phân tích cảm xúc và mô hình chủ đề. Phần tiếp theo trình bày về quy trình tạo ra bộ dữ liệu. Phần 4 gồm kết quả thực nghiệm trên bộ dữ liệu và phân tích lỗi trong các mô hình mà chúng tôi đã áp dụng. Kết luận và hướng phát triển của bài toán được giới thiệu trong phần 5. Và cuối cùng, phần 6 là các vấn đề về giới hạn và vấn đề đạo đức khi thực hiện bài toán.

II. Các nghiên cứu liên quan

A. Định nghĩa về bài toán nhận diện của cảm xúc và mô hình chủ đề

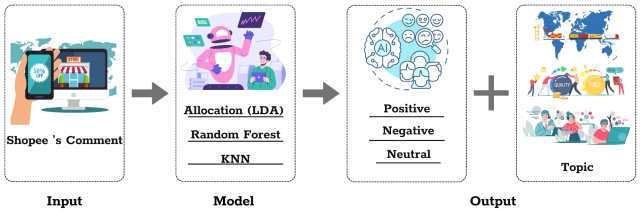
Input: Bình luận của khách hàng trên Shopee.

Output: Chủ đề tiềm ẩn và một trong ba nhãn tích cực, tiêu cực, trung lập được chúng tôi định nghĩa như sau:

[1]https://datareportal.com/reports/digital-2022-global-overview-report

[2]https://shopee.vn/

1) Nhãn tích cực: Nếu số lượng câu mang sắc thái tích cực trong bình luận chiếm hơn 50% tổng số câu thì bình luận đó mang sắc thái tích cực.
2) Nhãn tiêu cực: Nếu số lượng câu mang sắc thái tiêu cực trong bình luận chiếm hơn 50% tổng số câu thì bình luận đó mang sắc thái tiêu cực.
3) Nhãn trung lập: tất cả các câu mang sắc thái trung lập, 50% tổng số câu mang sắc thái tích cực và 50% còn lại mang sắc thái tiêu cực, câu vô nghĩa, câu mang sắc thái tích cực nhẹ hoặc tiêu cực nhẹ (có chứa từ "tạm").



Hình 1. Mô tả bài toán

B. Các nghiên cứu về phân tích cảm xúc trên thế giới

Maria Pontiki và các cộng sự [1] đã thực hiện các cuộc nghiên cứu về các khía cạnh của phân tích cảm xúc đối với các đánh giá thuộc bộ dữ liệu bao gồm: trích xuất chủ đề, phân loại chủ đề, phát hiện danh mục trong cùng một chủ đề, phân loại danh mục của chủ để. Nghiên cứu cho thấy kết quả của bài toán đối với từng khía cạnh là khác nhau. Bộ dữ liệu của Ganu [2] sử dụng sáu chủ đề (Food & Drink, Service, Price, Atmosphere, Anec dotes and Miscellaneous) và bốn nhãn đánh giá (positive, negative, conflict, neutral). McAuley và các cộng sự [3] cung cấp các khía cạnh và đánh giá về bia, đồ chơi, trò chơi, và sách nói. Các đánh giá được lấy từ các trang web cho phép người dùng đánh giá sản phẩm về cả mặt chất lượng cũng như các khía cạnh đặc trưng khác.

C. Các nghiên cứu về phân tích cảm xúc ở Việt Nam.

Ngo Xuan Bach và các cộng sự [4] đã thực hiện bài nghiên cứu phân tích cảm xúc dựa trên các khía cạnh khác nhau đối với bộ dữ liệu tiếng Việt, đây là phần mở rộng và phát triển từ bài toán của VLSP 2016. Luong Luc Phan [5] đã đạt được kết quả tốt nhất ($F1_{score}$) với 84.48% cho các khía cạnh của chủ đề và 63.06% cho phần phân tích cảm xúc thông qua máy học và các hệ thống học sâu.

D. Các nghiên cứu về mô hình chủ đề

Van-Ho Nguyen và các cộng sự [6] đã thực hiện một bài nghiên cứu về phân tích ý kiến khách hàng trực tuyến trong lĩnh vực khách sạn tiếp cận theo hướng mô hình chủ đề LDA. Kết quả cho thấy được tập chủ đề và các từ khóa trích xuất được đã phản ánh chính xác những vấn đề mà người dùng trong lĩnh vực khách sạn thường quan tâm.

III. Bộ dữ liệu

A. Thu thập dữ liệu

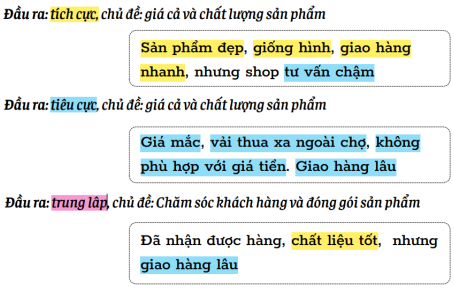
Để nhận diện cảm xúc của người tiêu dùng dựa trên các bình luận sau đó đưa ra chủ đề của bình luận, chúng tôi tiến hành lấy dữ liệu từ sàn thương mại điện tử Shopee thông qua Selenium và Beautifulsoup. Kết hợp BeautifulSoup [3] và Selenium[4] là một cách hiệu quả mà chúng tôi sử dụng để thu thập dữ liệu từ các trang web. BeautifulSoup giúp phân tích cú pháp và trích xuất thông tin từ mã HTML, trong khi Selenium cho phép chúng tôi tương tác với trang web động. Cuối cùng, chúng tôi thu thập được 7224 mẫu dữ liệu. Mỗi mẫu dữ liệu chứa bình luận và đánh giá sao của khách hàng sau khi trải nghiệm sản phẩm và các dịch vụ khác liên quan như: vận chuyển, chăm sóc khách hàng, chất lượng sản phẩm.

B. Quá trình gán nhãn

Trong quá trình gắn nhãn, chúng tôi sử dụng Fleiss's Kappa [7], một công cụ đo lường được sử dụng phổ biến để đánh giá sự đồng thuận giữa những người tham gia gán nhãn.

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Giai đoạn 1: Cả ba thành viên trong nhóm tiến hành huấn luyện gán nhãn, sau đó thông qua việc sử dụng Fleiss's kappa [7] để tính độ đồng thuận. Dựa trên các mẫu đã được gán nhãn, chúng tôi thảo luận và thống nhất để đưa ra nguyên tắc gắn nhãn chung cho cả bài toán và tiếp tục gán nhãn thêm 3000 dữ liệu bất kỳ trong bộ dữ liệu STiCS đã được thu thập trước đó. Quá trình này được lặp lại liên tục và độc lập dựa trên các quy tắc đã đặt ra cho đến khi độ đồng thuận giữa các thành viên cao hơn 80%.



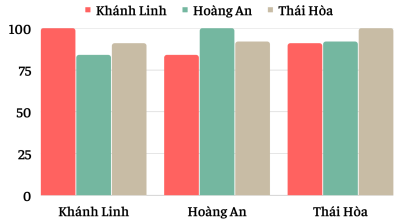
Hình 2. Một vài ví dụ khi gắn nhãn

Giai đoạn 2: Đánh ground Truth. Đầu tiên, ba thành viên trong nhóm đồng thời thực hiện quá trình gán nhãn lên 10% dữ liệu khác nhau của bộ dữ liệu STiCS, quá trình này được diễn ra một cách độc lập giữa ba người. Sau đó, thông qua mẫu dữ liệu đã được gán nhãn trên, chúng tôi sử dụng nó làm tập Ground Truth để tiến hành kiểm tra, đánh giá chéo nhau dựa trên việc tính toán độ đồng thuận bằng công thức Fleiss's Kappa [7].

Giai đoạn 3: Đánh nhãn chính thức. Chúng tôi tiến hành chia bộ dữ liệu thành ba phần, ứng với ba thành viên trong nhóm. Dựa trên các quy tắc gán nhãn ở phần trước, mỗi một thành

[3]https://pypi.org/project/beautifulsoup4/
[4]https://pypi.org/project/selenium/

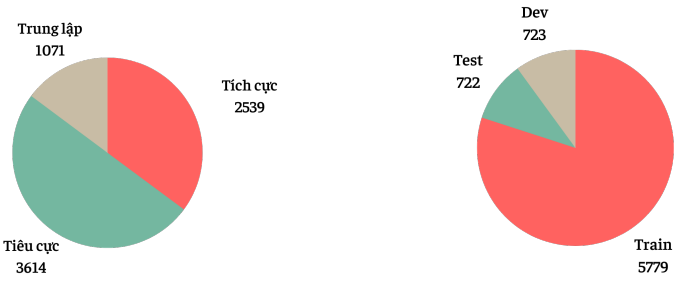
viên cần phải hoàn thành phần dữ liệu được phân công trong vòng bốn tuần. Để đảm bảo độ chính xác của bộ dữ liệu, mỗi một thành viên trong nhóm phải tự mình kiểm tra lại các đánh giá mà họ đã thực hiện và phản hồi kết quả lại với các thành viên còn lại trong nhóm. Ngoài ra, để giảm sai sót xuống mức tối thiểu, chúng tôi còn tiến hành kiểm tra 10% bình luận bất kì đã được gán nhãn của từng thành viên, so sánh với tập ground truth nếu tỉ lệ đồng thuận Fleiss's Kappa [7] thấp hơn 80%, thành viên đó phải thực hiện lại nhiệm vụ cho đến khi đạt được sự đồng thuận của cả nhóm.



Hình 3. Độ đồng thuận của 3 người gắn nhãn

C. Thống kê bộ dữ liệu

Chúng tôi tiến hành xử lý dữ liệu thô thành dữ liệu sẵn sàng, sau đó thực hiện gán nhãn trên bộ dữ liệu đã được xử lý. Bộ dữ liệu mới của chúng tôi bao gồm 7224 dòng và hai thuộc tính: comment và label. Trong đó, thuộc tính label gồm có ba nhãn: 0 tương ứng với tích cực (positive), 1 tương ứng với tiêu cực (negative), 2 tương ứng với trung lập (neutral). Sau khi hoàn tất quy trình gán nhãn, kết quả cho thấy bộ dữ liệu của chúng tôi có 2539 câu bình luận tích cực, 3614 bình luận tiêu cực và 1071 bình luận trung lập. Ngoài ra, sau khi tiến hành phân tích sơ bộ số lượng từ trong một câu bình luận chúng tôi thu được bảng VI.



Hình 4. Phân phối của dữ liệu

Chúng tôi tiến hành chia tập dữ liệu sẵn sàng thành ba tập huấn luyện (train), phát triển (dev) và kiểm thử (test) để thí nghiệm các mô hình. Ba tập dữ liệu được chia (theo phân phối nhãn - tức là phân phối nhãn trong cả ba tập là như nhau) theo tỉ lệ train: dev : test là 8 : 1 : 1.

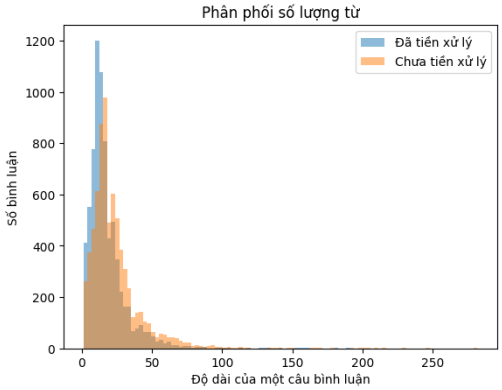
IV. Các phương pháp thực nghiệm

A. Tiền xử lý dữ liệu

Nhằm nâng cao hiệu quả của mô hình học máy chúng tôi thực hiện các phương pháp để xử lý bộ dữ liệu trước khi tiến hành

Bảng I
Thống kê bộ dữ liệu

Khía cạnh	Số lượng
Tổng số từ vựng	8538
Số từ nhiều nhất trong 1 bình luận	282
Số từ ít nhất trong 1 bình luận	1
Số từ trung bình của bình luận	23,36



Hình 5. Phân phối số lượng từ trong một bình luận

thực nghiệm. Các phương pháp xử lý bao gồm: (1) Xoá các ký tự không cần thiết (dấu câu, khoảng trắng thừa, icon). Ví dụ: sản phẩm rất đẹp ☺→ sản phẩm rất đẹp; (2) Chuyển các ký tự in hoa thành ký tự thường; (3) Chuẩn hóa tiếng Việt với bảng mã Unicode dựng sẵn; (4) Chuẩn hóa quy tắc đặt dấu câu trong tiếng Việt. Ví dụ: oà → òa; (5) Chuẩn hóa chữ viết kéo dài. Ví dụ: giao hàng nhanhhhhhhh → giao hàng nhanh; (6) Chuẩn hóa dữ liệu có chứa teencode; (7) Loại bỏ các kí tự đặc biệt và stopwords; (8) Phân tách theo từ (Word_Segmentation). Ví dụ: áo đẹp quá → 'áo', 'đẹp', 'quá'; (9) qua quá trình thống kê bộ dữ liệu, chúng tôi nhận thấy sự phân phối giữa các nhãn dự đoán là không đều, cụ thể nhãn chiếm đa số là nhãn tiêu cực, sự mất cân bằng này làm giảm đi độ chính xác của các mô hình máy học. Để cân bằng bộ dữ liệu, chúng tôi sử dụng phương pháp Oversampling cụ thể là SMOTE.

B. Các mô hình áp dụng thực nghiệm

1) Mô hình Naive Bayes [8]: Naive Bayes Classification là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes và thuộc nhóm học có giám sát. Theo định lý Bayes, ta có công thức tính xác suất ngẫu nhiên của sự kiện y khi biết x như sau: $P(x) = \frac{P(x)P(y)}{P(x)}$. Giả sử ta phân chia 1 sự kiện x thành n thành phần khác nhau x1, x2,..., X_n. Ta có:

$$P(x) \propto P(x) \prod_{i=1}^{n} P\left(x_i \mid y\right)$$

Mô hình này được áp dụng như một baseline model để so sánh với các mô hình khác giúp chúng tôi định lượng được hiệu năng

tối thiểu, dự kiến trong việc giải quyết bài toán phân loại ba nhãn.

2) Mô hình Support Vector Machines [9]: là một kỹ thuật phân lớp dữ liệu, là phương pháp học sử dụng không gian giả thuyết các hàm tuyến tính trên không gian đặc trưng nhiều chiều, dựa trên lý thuyết tối ưu và lý thuyết thống kê. Trong kỹ thuật SVM không gian dữ liệu nhập ban đầu sẽ được ánh xạ vào không gian đặc trưng và trong không gian đặc trưng này mặt siêu phẳng phân chia tối ưu sẽ được xác định. Siêu phẳng được biểu diễn bằng hàm số <W.X> =b (W và X là các vector; <W.X> là tích vô hướng) hay $W^T = b$ (W^T là ma trận chuyển vị). Trong phạm vi nghiên cứu, chúng tôi sử dụng hai biến thể SVC và LinearSVC.

3) Mô hình Logistic Regression: là một thuật toán machine learning có giám sát mạnh mẽ được sử dụng cho các bài toán phân loại nhị phân (khi mục tiêu là phân loại). Sự khác biệt cơ bản giữa hồi quy tuyến tính và hồi quy logistic là phạm vi của hồi quy logistic bị giới hạn từ 0 đến 1. Logistic sử dụng hàm Sigmoid:

$$FunctionLogistic = \frac{1}{1 + e^{-x}}$$

4) Mô hình kết hợp (Ensemble Learning): là một kỹ thuật học máy mà kết hợp nhiều mô hình học máy khác nhau để tạo ra một mô hình dự đoán có độ chính xác cao hơn so với các mô hình đơn lẻ. Phương pháp này thường được áp dụng trong các bài toán phân loại hoặc hồi quy, và đã được sử dụng rộng rãi trong các cuộc thi học máy. Có nhiều kỹ thuật Ensemble phổ biến như Bagging, Boosting và Stacking.

5) Non-negative Matrix Factorization (NMF) [10]: là một phương pháp giảm chiều dữ liệu dựa trên phân tích ma trận không âm. NMF tìm cách biểu diễn một ma trận không âm X (m x n) dưới dạng tích của hai ma trận không âm W (m x k) và H (k x n), với k nhỏ hơn m và n. NMF được ứng dụng trong Topic Modeling để phân tích ma trận tài liệu-thuật ngữ. Mỗi hàng của ma trận W biểu thị mức độ quan trọng của một chủ đề trong một tài liệu còn mỗi cột của ma trận H biểu thị mức độ quan trọng của một từ trong một chủ đề. Bằng cách tối thiểu hóa sai số giữa X và tích của W và H, NMF tìm ra các chủ đề ẩn trong dữ liệu văn bản.

6) Latent Dirichlet Allocation (LDA) [11]: là một mô hình xác suất chủ đề phổ biến, dựa trên phân phối xác suất Dirichlet. LDA giả định rằng mỗi tài liệu được tạo ra từ một phân phối xác suất của các chủ đề và mỗi chủ đề được tạo ra từ một phân phối xác suất của các từ. Trong LDA, chúng ta cố gắng tìm phân phối chủ đề-tài liệu (θ) và phân phối từ-chủ đề (ϕ) sao cho xác suất sinh dữ liệu được tối đa hóa. Để tìm các phân phối này, chúng ta sử dụng thuật toán suy diễn biến tiềm ẩn (Gibbs Sampling, Variational Inference) hoặc tối ưu hóa hợp lý.

C. Độ đo đánh giá

Accuracy: Độ chính xác tổng quát là tỷ lệ giữa số mẫu được phân loại đúng và tổng số mẫu trong tập dữ liệu. Độ chính xác tổng quát thể hiện khả năng phân loại chính xác của mô hình phân lớp trên toàn bộ tập dữ liệu, không phân biệt giữa các lớp.

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

$F1_{macro}$**:** một chỉ số đánh giá hiệu suất của mô hình phân lớp trong bài toán đa lớp. $F1_{score}$ là trung bình điều hòa giữa độ chính xác (precision) và độ đo bao hàm (recall), giúp đánh giá mô hình phân lớp một cách toàn diện hơn so với chỉ dùng độ chính xác hoặc độ đo bao hàm riêng lẻ.

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

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

$$F1_{score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$F1_{macro} = \frac{\sum_{i=1}^{n} F1_{scorei}}{n}$$

Topic coherence: được giới thiệu bởi Röder và các cộng sự [12] là chỉ số đánh giá chủ đề đơn lẻ dựa trên mức độ tương đồng ngữ nghĩa của các từ có trọng số cao trong chủ đề. Chỉ số này dùng phân biệt chủ đề có ý nghĩa về mặt ngữ nghĩa so với chủ đề được tạo ra từ kết quả của suy luận thống kê. Chỉ số Topic coherence có nhiều biến thể khác nhau, trong phạm vi bài báo cáo, chúng tôi sử dụng chỉ số C_v measure được tính dựa trên điểm NPMI-Normalized Pointwise Mutual Information và độ tương đồng Cosine.

Human Judgements: một phương pháp đánh giá chủ quan dựa trên sự đánh giá của con người. Sau khi tìm ra các mô hình có điểm số Topic Coherence cao, chúng tôi sử dụng trực giác cũng như sự hiểu biết của mình để lựa chọn mô hình có mức độ diễn giải tốt nhất, điều này nhằm đảm bảo các chủ đề được tạo ra thực sự có ý nghĩa và hữu ích trong thực tế.

D. Các thông số cài đặt mô hình

Mô hình Logistic Regression + Count Vectorizer: ngram_range = (1,3), min_df = 0.0004, max_df = 0.95, C = 12.2876, solver = 'lbfgs'. **Logistic Regression + TFIDF vectorizer:** ngram_range = (1,3), min_df = 0.0004, max_df = 0.95, C = 12.2876, solver = 'lbfgs'.

Mô hình SVC + Count Vectorizer: ngram_range = (1,3), min_df = 0.0004, max_df = 0.95, C = 10.9284, gamma = 1, kernel = 'rbf'. **SVC + TFIDF vectorizer:** ngram_range = (1,2), min_df = 0.0016, max_df = 0.95, C = 1.6475, gamma = 1, kernel = 'rbf'.

Mô hình LinearSVC + Count Vectorizer: ngram_range = (1,2), min_df = 0.0002, max_df = 0.95, C = 0.0679, loss = 'squared_hinge'. **LinearSVC + TFIDF vectorizer:** ngram_range = (1,2), min_df = 0.0001, max_df = 0.95, C = 1.1838, loss = 'hinge'.

Mô hình NMF + TF-IDF vectorizer: n_components = 5, max_df = 0.95, ngram_range = (1,1). **Mô hình LDA + TF-IDF vectorizer:** n_components = 4, max_df = 0.95, ngram_range = (1,1), n_jobs = -1.

E. Kết quả thực nghiệm

1) Đối với bài toán phân loại cảm xúc: Theo bảng II chúng tôi thấy rằng việc áp dụng phương pháp mã hóa TF-IDF vectorizer cho ra kết quả tốt hơn phương pháp count vectorizer trong hầu hết các thực nghiệm, việc áp dụng phương pháp SVMSMOTE trong việc cân bằng dữ liệu đã giúp tăng điểm số $F1_{Macro}$ trong các mô hình học máy mà chúng tôi đã huấn luyện. Sau khi chọn ra các mô hình với các bộ tham số tối ưu

Bảng II
KẾT QUẢ THỰC NGHIỆM CÁC MÔ HÌNH

Model	**Vectorize methods**	**Oversampling methods**	**Accuracy**	$F1_{macro}$
MultinomialNB	TF-IDF vectorizer	không áp dụng	72.12	53.02
	Count vectorizer	không áp dụng	**75.97**	**62.92**
LR	TF-IDF vectorizer	SVMSMOTE	79.19	**71.45**
		không áp dụng	**80.03**	71.06
	Count vectorizer	SVMSMOTE	76.60	69.49
		không áp dụng	78.95	67.04
SVC	TF-IDF vectorizer	SVMSMOTE	79.49	**70.11**
		không áp dụng	**79.78**	69.73
	Count vectorizer	SVMSMOTE	74.96	67.79
		không áp dụng	78.41	68.07
LinearSVC	TF-IDF vectorizer	SVMSMOTE	**80.03**	**71.04**
		không áp dụng	79.86	67.22
	Count vectorizer	SVMSMOTE	75.82	68.82
		không áp dụng	78.95	67.17

từ kết quả ở bảng trên, chúng tôi áp dụng phương pháp mô hình kết hợp để kết hợp chúng lại với nhau nhằm tăng hiệu suất của mô hình cuối cùng, kết quả thu được chúng tôi trình bày dưới đây:

Bảng III
KẾT QUẢ MÔ HÌNH ENSEMBLE

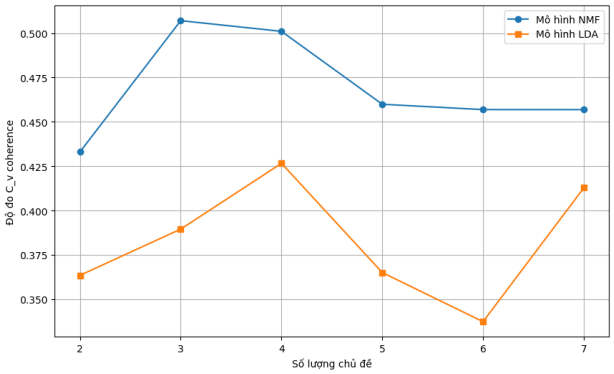
Ensemble model	**Tham số**	**Accuracy**	$F1_{macro}$
Voting	Soft	**80.60**	**71.34**
	Hard	80.44	71.26
Stacking	-	**80.84**	**71.60**

Mô hình Voting đối với độ đo Accuracy cho ra kết quả cao nhất là 80.60% và độ đo F1 macro cho ra kết quả cao nhất là 71.34% với tham số Soft và phương pháp mã hóa TFIDF. Mô hình Stacking với độ đo Accuracy cho ra kết quả là 80.84% và độ đo F1 macro cho ra kết quả là 71.60%. Nhìn chung, mô hình kết hợp cho ra kết quả tốt hơn so với các mô hình máy học riêng lẻ. Do vậy, chúng tôi quyết định dùng hai mô hình này để đánh giá kết quả trên tập test, kết quả được thể hiện ở bảng IV.

Bảng IV
KẾT QUẢ MÔ HÌNH TRÊN TẬP TEST

Model	Accuracy	$F1_{macro}$
Voting	80.93	70.63
Stacking	81.04	70.59

2) Đối với mô hình chủ đề: Chúng tôi nhận thấy mô hình NMF mang lại kết quả tốt hơn so với mô hình LDA. Số lượng chủ đề tối ưu cho mô hình NMF là 3, 4 và 5. Trong quá trình phân tích và đánh giá hiệu suất của các mô hình, chúng tôi đã phát hiện ra rằng mô hình NMF cho kết quả tốt hơn so với mô hình LDA. Để khảo sát kỹ hơn, chúng tôi đã tiến hành xem xét



Hình 6. So sánh độ đo C_v của LDA và NMF theo số lượng chủ đề

10 từ xuất hiện nhiều nhất trong mỗi chủ đề trong bảng V được tạo ra bởi mô hình NMF với số lượng chủ đề lần lượt là 3, 4 và 5. Kết quả cho thấy mô hình NMF với số lượng chủ đề là 5 cho ra kết quả có ý nghĩa nhất. Dựa trên top 10 từ xuất hiện nhiều nhất của mỗi chủ đề, chúng tôi đã đặt tên cho các chủ đề sao cho phù hợp với lĩnh vực thương mại điện tử được đề cập trong bài báo cáo này.

Bảng V
KẾT QUẢ CỦA MÔ HÌNH CHỦ ĐỀ

Chủ đề	**10 từ phổ biến trong chủ đề**	**Tên chủ đề**
1	hàng, shop, lẫn, chất lượng, người, đóng gói, đơn, cẩn thận, bên, ốp	Chăm sóc khách hàng và đóng gói sản phẩm
2	áo, vải, người, chất, hình, size, quần, lẫn, thừa, chất liệu	Chất lượng quần áo
3	tính chất, hình ảnh, xu, người, minh họa, ảnh, ok, tính, mô tả, đánh giá	Hình ảnh và minh họa sản phẩm
4	màu, shop, quần, hình, ảnh, ốp, lẫn, hồng, mẫu, ngoài	Màu sắc và hình ảnh sản phẩm
5	giá, tiền, sản phẩm, chất lượng, người , ốp, tầm, ok, lẫn, nói chung	Giá cả và chất lượng sản phẩm

F. Phân tích lỗi

1) Phân tích cảm xúc: Một trong những lỗi phổ biến là một vài câu không rõ sắc thái, gây khó khăn trong quá trình gán nhãn và ảnh hưởng hiệu suất model. Điều này xảy ra do một số câu chứa biểu cảm không rõ ràng, sắc thái không thể hiện đúng mức độ cảm xúc của người nói. Điều này dẫn đến việc gán nhãn không chính xác, ảnh hưởng đến sự học tập của mô hình và làm giảm hiệu suất phân loại. Ngoài ra, cũng có một vài câu không rõ nghĩa, chứa nhiều từ khó chuẩn hóa. Những câu này thường xuất hiện do việc sử dụng ngôn ngữ không chuẩn mực, viết tắt hoặc từ ngữ không thông dụng. Điều này gây khó khăn cho mô

hình trong việc học và hiểu được ý nghĩa của từng từ, dẫn đến việc phân loại cảm xúc không chính xác.

Điển hình là trường hợp sau: "hàng tốt nhưng thích số 2, rate 2* nhé sốp". Nhận xét này cho thấy, người mua rất hài lòng về sản phẩm (tích cực), nhưng do các nguyên nhân ngoại lai khiến câu trở thành tiêu cực, kết quả của việc gán nhãn là tiêu cực. Mặt khác, sự tồn tại của các bình luận không rõ nghĩa, vô nghĩa như: "aajfksdvsldkvsldkcsl" hoặc "hihihihihi" cũng sẽ ảnh hưởng đến kết quả dự đoán của mô hình.

Thông qua việc quan sát các dự đoán không chính xác trong bộ dữ liệu, chúng tôi đã phát hiện ra một số bình luận có chứa từ mang tính tích cực, nhưng trong một ngữ cảnh khác lại mang tính tiêu cực: "Áo rộng, mặc không lộ khuyết điểm". Có thể thấy, khi đứng một mình, từ "rộng" sẽ mang ý nghĩa tiêu cực, không phù hợp với cơ thể. Nhưng nhìn tổng thể cả câu, thì đây là một bình luận thể hiện sự hài lòng của khách hàng.

2) Mô hình chủ đề: Mô hình bị hạn chế trong việc học từ đặc trưng do kích thước bộ dữ liệu không đủ lớn, dẫn đến giảm khả năng diễn giải các chủ đề. Khi đánh giá các chủ đề được mô hình tạo ra, chúng tôi phát hiện một số chủ đề có khả năng diễn giải chưa đạt yêu cầu, nguyên nhân chủ yếu là các từ đặc trưng trong bộ dữ liệu không mang lại ý nghĩa rõ ràng. Thêm vào đó, việc lựa chọn số chủ đề phù hợp cho mô hình NMF không có tiêu chuẩn rõ ràng, điều này khiến việc huấn luyện và đánh giá mô hình trở nên thách thức.

V. Kết luận và hướng đi trong tương lai

Trong nghiên cứu lần này, chúng tôi đã áp dụng thành công các mô hình học máy để nhận diện được cảm xúc của khách hàng và đưa ra được các chủ đề tương ứng của bình luận dựa trên bộ dữ liệu STiSC do chúng tôi tự xây dựng. Đối với bài toán nhận diện cảm xúc, mô hình kết hợp của ba mô hình SVC, LinearSVC và Logistics Regression cho ra kết quả tốt nhất với 80,84% với độ đo accuracy và 71,6% với độ đo $F1_{macro}$, bên cạnh đó với bài toán mô hình chủ đề, mô hình NMF cho ra kết quả tốt nhất với 5 chủ đề.

Ở bài toán này, chúng tôi còn gặp một số khó khăn, khiến cho kết quả chưa đạt được như mong muốn do đó chúng tôi có những định hướng tiếp theo trong tương lai đối với nghiên cứu này như sau: mở rộng bộ dữ liệu, thử nghiệm các phương pháp tiền xử lý mới, thử nghiệm trên các mô hình học sâu, học chuyển tiếp kết hợp tinh chỉnh các mô hình và mở rộng đa dạng ngành hàng.

VI. Các giới hạn và vấn đề đạo đức

Một nhược điểm lớn trong quá trình thu thập dữ liệu từ Shopee là dữ liệu sau khi thu thập rất lộn xộn (messy data) thêm vào đó, dữ liệu văn bản lại là dữ liệu không cấu trúc, những điều này làm hao tổn một lượng chi phí không nhỏ để chúng tôi lên kế hoạch để xử lý. Liên quan đến các chính sách về quyền riêng tư và quyền sở hữu trí tuệ, chúng tôi cam đoan rằng mọi dữ liệu chúng tôi thu thập không chứa bất kỳ thông tin nhạy cảm nào của khách hàng và chúng tôi chỉ sử dụng nó cho mục đích nghiên cứu. Trong quá trình gán nhãn bình luận cảm xúc, người gán nhãn thường phải đọc và xử lý các bình luận có tính chất tiêu cực, lời lẽ tục tĩu, hay thậm chí là những bình luận xúc phạm. Việc tiếp xúc liên tục với những nội dung tiêu cực này có thể ảnh hưởng đến tâm trạng và sức khỏe tinh thần của người gán nhãn theo tạp chí The Verge[5].

Lời cảm ơn

Cảm ơn thầy Dương Ngọc Hảo và cô Nguyễn Lưu Thuỳ Ngân đã hỗ trợ nhiệt tình và giải đáp các thắc mắc của nhóm để thực hiện được đề tài này.

Tài liệu

[1] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah, "Aspect-based sentiment analysis: A survey of deep learning methods," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 6, pp. 1358–1375, 2020.

[2] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, Jun. 2010, pp. 804–812. [Online]. Available: https://aclanthology.org/N10-1122

[3] J. McAuley, J. Leskovec, and D. Jurafsky, "Learning attitudes and attributes from multi-aspect reviews," in *2012 IEEE 12th International Conference on Data Mining*, 2012, pp. 1020–1025.

[4] H. T. M. Nguyen, H. V. Nguyen, Q. T. Ngo, L. X. Vu, V. M. Tran, B. X. Ngo, and C. A. Le, "Vlsp shared task: Sentiment analysis," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, p. 295–310, Jan. 2019. [Online]. Available: https://vjs.ac.vn/index.php/jcc/article/view/13160

[5] L. Luc Phan, P. Huynh Pham, K. Thi-Thanh Nguyen, S. Khai Huynh, T. Thi Nguyen, L. Thanh Nguyen, T. Van Huynh, and K. Van Nguyen, "Sa2sl: From aspect-based sentiment analysis to social listening system for business intelligence," in *Knowledge Science, Engineering and Management*, H. Qiu, C. Zhang, Z. Fei, M. Qiu, and S.-Y. Kung, Eds. Cham: Springer International Publishing, 2021, pp. 647–658.

[6] N. Ho and H. Thanh, "Topic modeling for analyzing online reviews in hotel sector," *VNUHCM Journal of Economics, Business and Law*, vol. 4, no. 4, pp. 1081–1092, Nov. 2020. [Online]. Available: http://stdjelm.scienceandtechnology.com.vn/index.php/stdjelm/article/view/692

[7] D. G. Seigel, M. J. Podgo, and N. A. Remaley, "Acceptable Values of Kappa for Comparison of Two Groups," *American Journal of Epidemiology*, vol. 135, no. 5, pp. 571–578, 03 1992. [Online]. Available: https://doi.org/10.1093/oxfordjournals.aje.a116324

[8] G. I. Webb, *Naïve Bayes*. Boston, MA: Springer US, 2010, pp. 713–714. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_576

[9] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," vol. 2049, 09 2001, pp. 249–257.

[10] nov 2022. [Online]. Available: https://doi.org/10.1007%2Fs41060-022-00370-9

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, mar 2003.

[12] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. [Online]. Available: https://doi.org/10.1145/2684822.2685324

Phân công nhiệm vụ

Công việc của nhóm chúng em được phân chia như bảng sau:

[5]https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Bảng VI
PHÂN CÔNG NHIỆM VỤ

Công việc	**Thành viên thực hiện**
Viết Guideline	Hoàng An Thái Hoà
Craw dữ liệu	Hoàng An
Tiền xử lý dữ liệu	Hoàng An Khánh Linh Thái Hoà
Gán nhãn bộ dữ liệu	Khánh Linh Hoàng An THái Hoà
Kiểm thử mô hình	Hoàng An Thái Hoà
Thực nghiệm	Hoàng An Khánh Linh Thái Hoà
Viết báo cáo	Khánh Linh Hoàng An Thái Hoà
Slide	Khánh Linh