**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

****

**BÁO CÁO CUỐI KỲ MÔN HỌC**

**TÊN MÔN HỌC: PHÂN TÍCH DỮ LIỆU VỚI R/PYTHON**

**MÃ HỌC PHẦN: 232IS2901**

**“ĐỀ XUẤT PHÂN LOẠI VÀ ĐIỂM ĐÁNH GIÁ CHO BÌNH LUẬN CỦA NGÀNH HÀNG THỜI TRANG**

**TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ SHOPEE”**

**Giảng viên hướng dẫn: ThS Nguyễn Phát Đạt**

Danh sách thành viên nhóm:

|  |  |
| --- | --- |
| **MSSV** | **Họ và tên** |
| K214110827 | Nguyễn Anh Tuấn |
| K214111322 | Nguyễn Ngọc Phương Uyên |
| K214110786 | Nguyễn Hồng Ngọc Ánh |
| K214111952 | Trần Anh Tâm |
| K214110823 | Bùi Thị Anh Thuy |
| K214111320 | Phạm Anh Thư |

*Thành phố Hồ Chí Minh, tháng 05 năm 2024*

**MỤC LỤC**

[LỜI CẢM ƠN](#_Toc166002977)

[DANH MỤC BẢNG BIỂU](#_Toc166002978)

[DANH MỤC HÌNH ẢNH](#_Toc166002979)

[Chương 1. Giới thiệu 1](#_Toc166002980)

[1.1. Lý do lựa chọn đề tài 1](#_Toc166002981)

[1.2. Mục tiêu nghiên cứu 3](#_Toc166002982)

[1.2.1. Mục tiêu 3](#_Toc166002983)

[1.2.2. Mục đích 3](#_Toc166002984)

[1.3. Đối tượng và phạm vi nghiên cứu 4](#_Toc166002985)

[1.3.1. Đối tượng 4](#_Toc166002986)

[1.3.2. Phạm vi nghiên cứu 4](#_Toc166002987)

[1.4. Phương pháp nghiên cứu 4](#_Toc166002988)

[1.5. Ý nghĩa nghiên cứu 4](#_Toc166002989)

[1.5.1. Ý nghĩa khoa học 4](#_Toc166002990)

[1.5.2. Ý nghĩa thực tiễn 5](#_Toc166002991)

[1.6. Cấu trúc nghiên cứu 5](#_Toc166002992)

[Chương 2. Cơ sở lý thuyết và các nghiên cứu liên quan 6](#_Toc166002993)

[2.1. Các nghiên cứu liên quan 6](#_Toc166002994)

[2.1.1. Nghiên cứu quốc tế 6](#_Toc166002995)

[2.1.2. Nghiên cứu trong nước 7](#_Toc166002996)

[2.2. Cơ sở lý thuyết 8](#_Toc166002997)

[2.2.1. Phân tích cảm xúc 8](#_Toc166002998)

[2.2.2. API 8](#_Toc166002999)

[2.2.3. Fleiss’ Kappa 9](#_Toc166003000)

[2.2.4. Kỹ thuật nhúng từ - BERT Embedding 10](#_Toc166003001)

[2.2.5. K-Fold 10](#_Toc166003002)

[2.2.6. Máy học truyền thống 11](#_Toc166003003)

[2.2.7. Deep Learning: CNN 14](#_Toc166003004)

[2.2.8. Mô hình Transformer 14](#_Toc166003005)

[Chương 3. Phương pháp nghiên cứu 15](#_Toc166003006)

[3.1. Mô tả bài toán 15](#_Toc166003007)

[3.2. Đề xuất giải pháp 16](#_Toc166003008)

[3.2.1. Giải pháp 16](#_Toc166003009)

[3.2.2. Xây dựng mô hình giải pháp 17](#_Toc166003010)

[Chương 4. Kết quả nghiên cứu và thảo luận 18](#_Toc166003011)

[4.1. Mô tả cách lấy dữ liệu 18](#_Toc166003012)

[4.2. Tiền xử lý dữ liệu 22](#_Toc166003013)

[4.3. EDA 30](#_Toc166003014)

[4.4. Chuẩn bị mô hình 38](#_Toc166003015)

[4.4.1. Machine Learning 40](#_Toc166003016)

[4.4.2. Deep Learning 42](#_Toc166003017)

[4.4.3. Transformer 43](#_Toc166003018)

[4.5. Kết quả phân tích 45](#_Toc166003019)

[4.5.1. In-Out Domain 45](#_Toc166003020)

[4.5.1.1. No Token 45](#_Toc166003021)

[4.5.1.2. Token 46](#_Toc166003022)

[4.5.2. Predict Star 48](#_Toc166003023)

[4.5.2.1. No Token 48](#_Toc166003024)

[4.5.2.2. Token 49](#_Toc166003025)

[4.6. Ứng dụng mô hình 51](#_Toc166003026)

[Chương 5. Kết luận và hướng phát triển 53](#_Toc166003027)

[5.1. Tóm tắt nội dung và kết quả nghiên cứu của đề tài 53](#_Toc166003028)

[5.2. Ưu điểm 54](#_Toc166003029)

[5.3. Nhược điểm 55](#_Toc166003030)

[5.4. Hướng phát triển của đề tài 55](#_Toc166003031)

[Tài liệu tham khảo 56](#_Toc166003032)

[PHỤ LỤC 60](#_Toc166003033)

# LỜI CẢM ƠN

Để hoàn thành bài báo cáo này, nhóm 3 xin gửi lời cảm ơn chân thành đến Thầy Nguyễn Phát Đạt là giảng viên phụ trách môn học “Phân tích dữ liệu với R-Python”. Chúng em xin chân thành cảm ơn thầy vì đã giảng dạy và truyền đạt những kiến thức bổ ích, từ đó nhóm có thể hoàn thành những phần báo cáo này.

Nhóm đã rất cố gắng để tìm hiểu các nguồn tài liệu, nghiên cứu thêm nhiều thông tin

và học hỏi. Tuy nhiên, vì là lần đầu thực hiện đồ án môn với vốn kiến thức vẫn còn hạn

chế nên sẽ không thể tránh khỏi sai sót. Kính mong nhận được sự đóng góp ý kiến của thầy để báo cáo được hoàn thiện.

Một lần nữa, nhóm xin chân thành cảm ơn.

# DANH MỤC BẢNG BIỂU

[Bảng 1.1. Tình hình mua sắm trực tuyến tại việt nam năm 2019 - 2023 2](#_Toc166002970)

[Bảng 2.1. Cơ chế hoạt động của API 9](#_Toc166002971)

[Bảng 4.1. Bảng thông số trong mô hình phoBert 45](#_Toc166002972)

[Bảng 4.2. Kết quả huấn luyện in-out domain trên tập dữ liệu no-token 45](#_Toc166002973)

[Bảng 4.3. Kết quả huấn luyện in-out domain trên tập dữ liệu token 47](#_Toc166002974)

[Bảng 4.4. Kết quả huấn luyện dự đoán sao (predict star) trên tập dữ liệu no-token 48](#_Toc166002975)

[Bảng 4.5. Kết quả huấn luyện dự đoán sao (predict star) trên tập dữ liệu token 49](#_Toc166002976)

# DANH MỤC HÌNH ẢNH

[Hình 3.1. Mô hình giải pháp 17](#_Toc166002922)

[Hình 4.1. Xác định thông tin cơ bản của sản phẩm cần trích xuất 19](#_Toc166002923)

[Hình 4.2. Xác định thông tin về bình luận và đánh giá 19](#_Toc166002924)

[Hình 4.3. Lấy API sản phẩm 20](#_Toc166002925)

[Hình 4.4. Lấy API đánh giá và bình luận 21](#_Toc166002926)

[Hình 4.5. Vòng lặp while lấy tối đa 100 đánh giá và bình luận với mỗi sao 22](#_Toc166002927)

[Hình 4.6. Lưu dữ liệu thành tệp CSV 22](#_Toc166002928)

[Hình 4.7. Import thư viện emot 22](#_Toc166002929)

[Hình 4.8. Import thư viện py\_vncorenlp 23](#_Toc166002930)

[Hình 4.9. Import thư viện pickle 23](#_Toc166002931)

[Hình 4.10. Chuyển đổi dữ liệu thành DataFrame 24](#_Toc166002932)

[Hình 4.11. Dữ liệu gốc thu được sau khi lấy dữ liệu 24](#_Toc166002933)

[Hình 4.12. Thực hiện loại bỏ các bình luận bị rỗng hoặc lặp lại 25](#_Toc166002934)

[Hình 4.13. Bộ dữ liệu sau khi đã loại bỏ bình luận bị rỗng và lặp lại 25](#_Toc166002935)

[Hình 4.14. Tạo từ điển ánh xạ các từ ngữ viết tắt hoặc từ lóng (teencode) 26](#_Toc166002936)

[Hình 4.15. Cập nhật các từ viết tắt, từ lóng vào trong từ điển đã tạo 26](#_Toc166002937)

[Hình 4.16. Kết quả số lượng từ viết tắt và từ lóng (teencode) trong từ điển 27](#_Toc166002938)

[Hình 4.17. Cập nhật từ điển từ tệp văn bản khác 27](#_Toc166002939)

[Hình 4.18. Xây dựng từ điển về kích thước của mặt hàng thời trang 27](#_Toc166002940)

[Hình 4.19. Chuyển đổi emoji và emoticon thành văn bản 28](#_Toc166002941)

[Hình 4.20. Chuẩn hoá dữ liệu văn bản của bình luận (1) 28](#_Toc166002942)

[Hình 4.21. Chuẩn hoá dữ liệu văn bản của bình luận (2) 29](#_Toc166002943)

[Hình 4.22. Dữ liệu bình luận sau khi đã được chuẩn hoá và làm sạch 29](#_Toc166002944)

[Hình 4.23. Thêm bình luận đã được xử lý và làm sạch vào DataFrame 30](#_Toc166002945)

[Hình 4.24. Bộ dữ liệu thu được sau khi đã thực hiện quá trình tiền xử lý 30](#_Toc166002946)

[Hình 4.25. Thư viện sử dụng trong quá trình EDA 30](#_Toc166002947)

[Hình 4.26. Dữ liệu được sử dụng để khai phá dữ liệu 31](#_Toc166002948)

[Hình 4.27. Thông tin bộ dữ liệu 32](#_Toc166002949)

[Hình 4.28. Mô tả cột dữ liệu rating star 32](#_Toc166002950)

[Hình 4.29. Đếm số lượng từng sao trong cột ‘rating\_star’ và sự phân bổ 33](#_Toc166002951)

[Hình 4.30. Hàm xử lý xoá emoji có định dạng <emoji> </emoji> 33](#_Toc166002952)

[Hình 4.31. Xoá những dữ liệu mặc định có sẵn trong phần đánh giá 34](#_Toc166002953)

[Hình 4.32. Biểu đồ histogram thể hiện số lượng ký tự có trong từng đánh giá 34](#_Toc166002954)

[Hình 4.33. Số lượng ký tự trung bình 34](#_Toc166002955)

[Hình 4.34. Biểu đồ histogram thể hiện số từ có trong từng đánh giá 35](#_Toc166002956)

[Hình 4.35. Sự phân bổ số lượng đánh giá có trên 100 từ và dưới 100 từ 35](#_Toc166002957)

[Hình 4.36. Tokenize và tạo trigram cho mỗi đánh giá 36](#_Toc166002958)

[Hình 4.37. Biểu đồ thể hiện tần suất xuất hiện của 10 triagram phổ biến nhất 36](#_Toc166002959)

[Hình 4.38. WordCloud 37](#_Toc166002960)

[Hình 4.39. Biểu thị sự tương quan giữa ‘rating\_star' và ‘clean\_cmt' 38](#_Toc166002961)

[Hình 4.40. Gắn nhãn 1000 dòng dữ liệu đầu tiên 39](#_Toc166002962)

[Hình 4.41. Sử dụng Fleiss Kappa để đo độ chính xác gắn nhãn dữ liệu 40](#_Toc166002963)

[Hình 4.42. Chuẩn bị hàm Machine Learning 42](#_Toc166002964)

[Hình 4.43. Chuẩn bị hàm Deep Learning 43](#_Toc166002965)

[Hình 4.44. Chuẩn bị hàm Transformer 44](#_Toc166002966)

[Hình 4.45. Kết quả ứng dụng mô hình 51](#_Toc166002967)

[Hình 4.46. Mockup ứng dụng để đề xuất phân loại và điểm đánh giá 52](#_Toc166002968)

[Hình 4.47. Dashboard đánh giá của sản phẩm 53](#_Toc166002969)

# Giới thiệu

## Lý do lựa chọn đề tài

Trong bối cảnh công nghệ ngày càng phát triển, nhu cầu của con người trong cuộc sống cũng ngày càng đổi mới và nâng cấp hơn. Sau đại dịch Covid-19 là sự chuyển dịch mạnh mẽ từ thói quen mua sắp trực tiếp sang mua sắm trực tuyến của người tiêu dùng.

Theo báo cáo của iDEA (Cục Thương mại điện tử và Kinh tế số Việt Nam), thị trường mua sắm trực tuyến tại Việt Nam đang phát triển mạnh mẽ, với mức tăng trưởng ước tính lên đến 25-30% mỗi năm.

Xu hướng mua sắm của mọi người dần thay đổi, họ không còn phải tốn nhiều thời gian để chọn lựa mua sắm tại các cửa hàng thời trang vật lý, chỉ cần có chiếc điện thoại thông minh có kết nối internet, mọi người có thể mua sắm tại bất cứ đâu và bất kỳ thời điểm nào. Nếu như trước đây chúng ta phải dành nhiều thời gian đi nhiều cửa hàng để lựa chọn một bộ quần áo ưng ý, thì ngày nay việc mua sắm online đã giúp điều đó trở nên tiện lợi hơn bao giờ hết. Những lý do đó đã khiến ngành hàng thời trang trên các sàn thương mại điện tử có doanh số vượt trội và ngày càng phát triển mạnh mẽ. Tốc độ phát triển thương mại điện tử tại Việt Nam năm 2022 đạt hơn 20%. Dự kiến, đến năm 2025 con số này có thể lên đến 29%, với giá trị khoảng 234 tỷ USD. Hiện nay cuộc chơi trên thị trường TMĐT trong nước rơi vào tay 5 sàn thương mại lớn nhất là Shopee, Lazada, Tiki, Sendo và Tiktok Shop. Trong đó, hiện tổng doanh thu trên 5 sàn thương mại bán lẻ trực tuyến lớn nhất hiện nay đã đạt mức 232,134 tỷ đồng, tăng trưởng 53.4% so với cùng kỳ năm 2022, với tổng số sản phẩm giao dịch thành công lên tới 2.2 tỷ đơn vị.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **2019** | **2020** | **2021** | **2022** | **2023** |
| Số liệu người tiêu dùng mua sắp trực tuyến (triệu người) | 44.8 | 49.3 | 54.6 | 57.0 | 61.0 |
| Ước tính giá trị mua sắm trực tuyến của 1 người (USD) | 225 | 240 | 251 | 288 | 336 |
| Tỷ trọng doanh thu TMĐT B2C so với tổng mức bán lẻ hàng hoá và doanh thu dịch vụ tiêu dùng cả nước | 4.9% | 5.5% | 7.0% | 7.5% | 8.0% |
| Tỷ lệ người dân sử dụng Internet | 66.0% | 70.0% | 73.0% | 73.2% | 78.6% |

Bảng 1.1. Tình hình mua sắm trực tuyến tại việt nam năm 2019 - 2023

Theo số liệu thống kê từ Metric, chỉ tính từ 10-8-2023 đến khoảng giữa tháng 9, doanh số tổng cộng trên sàn Shopee là khoảng 14.000 tỉ đồng. Trong đó, ngành hàng sắc đẹp dẫn đầu với doanh số khoảng 2.400 tỷ đồng, tiếp theo là danh mục nhà cửa và đời sống với khoảng 1.800 tỷ đồng, và ngành thời trang nữ đứng ở vị trí thứ ba với khoảng 1.500 tỷ đồng. Trong khi đó, doanh số của ngành thời trang nam ước tính chỉ là 500 tỷ đồng, xếp gần cuối trong danh sách. Điều này phản ánh sự phát triển của ngành thời trang trong thương mại điện tử.

Có thể nhận thấy rằng, đây chính là một ngành hàng vô cùng tiềm năng, đặc biệt là trên nền tảng Shopee, khi mà họ đã kết nối được với vô vàn nhà cung cấp thời trang trong và ngoài nước. Tuy vậy, nhu cầu của khách hàng ngày càng nâng cao, đòi hỏi các doanh nghiệp kinh doanh cũng phải thay đổi theo thị trường. Do đó, việc hiểu rõ được mức độ hài lòng cũng như là sở thích của khách hàng trở nên quan trọng hơn bao giờ hết. Bởi nó sẽ là chìa khóa đáng giá giúp cho các doanh nghiệp có thể cải thiện chất lượng, cá nhân hóa trải nghiệm mua sắm, cung cấp sản phẩm và dịch vụ phù hợp với mong đợi của từng người tiêu dùng.

Nhóm chúng em chọn đề tài **“Đề xuất phân loại và điểm đánh giá cho bình luận của ngành hàng thời trang trên sàn thương mại điện tử Shopee”** với mong muốn từ việc áp dụng học máy để phân tích dữ liệu, từ đó hiểu được mức độ hài lòng cũng như là sở thích người tiêu dùng từ những phản hồi của họ, giúp cho doanh nghiệp kinh doanh thời trang trên Shopee tối ưu hoá quy trình sản xuất và quảng cáo sản phẩm, cũng như là nâng cao trải nghiệm mua sắm của khách hàng.

**Tính cấp thiết của đề tài**

Hiện nay, nghiên cứu và đánh giá về cảm xúc khách hàng luôn là yếu tố quan trọng tạo nên lợi thế cạnh tranh cho doanh nghiệp. Internet và smartphone đóng vai trò quan trọng trong việc kết nối khách hàng với doanh nghiệp. Nhờ đó, khách hàng có thể dễ dàng tìm kiếm thông tin, mua sắm sản phẩm, dịch vụ và chia sẻ ý kiến của mình về sản phẩm, dịch vụ hoặc thương hiệu một cách trực tiếp. Rào cản lớn nhất chính là khối lượng dữ liệu bình luận khổng lồ, dẫn đến việc tiếp cận và nắm bắt đầy đủ thông tin từ tất cả các bình luận là một vấn đề nan giải.

Nhằm thấu hiểu khách hàng và xây dựng chiến lược kinh doanh hiệu quả, việc ứng dụng công nghệ mới là vô cùng cần thiết. Các phương pháp học máy sẽ phân tích cảm xúc bình luận và đánh giá độ tin cậy của bình luận, nhờ đó mà doanh nghiệp có thể nắm bắt được ý kiến, quan điểm của khách hàng về sản phẩm, từ đó hiểu rõ thói quen và hành vi mua sắm của họ. Dựa vào đánh giá của khách hàng doanh nghiệp có thể xác định điểm không hài lòng từ đó có thể đề xuất các phương án cải thiện chất lượng sản phẩm, ngoài ra còn có thể có các phương án cái thiện các vấn đề liên quan đến dịch vụ giao hàng, chăm sóc khách hàng của doanh nghiệp.

Nhìn chung, việc đánh giá ý kiến phản hồi của khách hàng thông qua các công nghệ trí tuệ nhân tạo mới sẽ giúp đo lường cảm xúc của khách hàng về sản phẩm trên thị trường một cách hiệu quả. Điều này giúp các đơn vị cung cấp sản phẩm hiểu rõ hơn về khách hàng, tối ưu hóa quá trình thu thập thông tin và đưa ra các quyết định chính xác hơn.

## Mục tiêu nghiên cứu

### Mục tiêu

Mục tiêu của đề tài là thông qua đặc điểm đánh giá của khách hàng để có thể phân loại đánh giá, từ đó phân tích mức độ hài lòng của khách hàng trên sàn thương mại điện tử Shopee nhằm tạo ra một công cụ hỗ trợ đánh giá độ đáng tin cậy của thông tin phản hồi từ người tiêu dùng trên nền tảng này. Thu thập dữ liệu để phân tích và thống kê nhằm cung cấp cho doanh nghiệp tài liệu tham khảo cho các chiến lược tiếp thị, thiết lập lại nhu cầu và dịch vụ của khách hàng. Đồng thời, đề tài sẽ xây dựng một giao diện hệ thống phân loại và đề xuất điểm đánh giábình luận giúp doanh nghiệp dễ dàng nhận biết được tổng quan về những lượt đánh giá sản phẩm của mình.

### Mục đích

Đề tài được nghiên cứu và phát triển đáp ứng theo các yêu cầu dưới đây:

* Phát triển một mô hình dự đoán dựa trên Machine Learning, Deep Learning, Transformer để đánh giá độ chính xác của điểm đánh giá của bình luận sản phẩm trên sàn thương mại điện tử Shopee.
* Đánh giá hiệu suất của bộ máy dự đoán, từ đó đề xuất cách thức áp dựng bộ máy dự đoán vào thực tế
* Đánh giá tác động của bình luận đến quyết định mua sắm trực tuyến và lòng tin của người tiêu dùng đối với sản phẩm.
* Phân tích bình luận và đánh giá của sản phẩm để xem xét về mức độ chân thật trong việc đánh giá.
* Tạo ra nhận thức về độ trung thực của sản phẩm trên Shopee, giúp người tiêu dùng có quyết định mua sắm dựa trên thông tin đánh giá sản phẩm một cách chính xác và đáng tin cậy.

## Đối tượng và phạm vi nghiên cứu

### Đối tượng

Cảm xúc trong bình luận của từng khách hàng trên sàn thương mại điện tử Shopee ngành hàng Thời trang

### Phạm vi nghiên cứu

* Phạm vi không gian: Đánh giá và bình luận trên sàn thương mại điện tử Shopee

ngành hàng Thời trang

* Phạm vi thời gian: Tháng 12/2023 - 02/2024

## Phương pháp nghiên cứu

* Phương pháp định tính: Dựa trên việc nghiên cứu phân tích các công trình nghiên cứu liên quan, từ đó đánh giá ưu nhược điểm của các hướng tiếp cận và đề xuất phương pháp phù hợp với ứng dụng cụ thể.
* Nghiên cứu thực nghiệm: Từ các phương pháp đề xuất, sau khi đã được phân tích đánh giá định tính sẽ được thử nghiệm và đánh giá trên các bộ dữ liệu thực tế.

## Ý nghĩa nghiên cứu

### Ý nghĩa khoa học

Bài nghiên cứu sẽ đề xuất được mô hình nghiên cứu có tính ứng dụng cao trong việc đề xuất phân loại và điểm đánh giá cho bình luận cho ngành hàng thời trang của khách hàng khi mua sắm trên sàn thương mại điện tử. Mô hình này có thể được ứng dụng để phân tích dữ liệu và đưa ra dự đoán cho những ngành hàng kinh doanh trực tuyến tương tự.

### Ý nghĩa thực tiễn

Đề tài nghiên cứu do nhóm đề xuất có ý nghĩa quan trọng trong thực tiễn, giúp người đọc có cái nhìn sâu sắc hơn về ngành hàng thời trang trực tuyến, đồng thời giúp khách hàng đưa ra đánh giá khách quan hơn khi đánh giá các sản phẩm thời trang, từ đó giúp doanh nghiệp kinh doanh trực tuyến cải thiện tình hình kinh doanh và các dịch vụ liên quan đến sản phẩm, đặc biệt là các cửa hàng thời trang trên sàn thương mại điện tử Shopee. Những nhà quản lý cũng có thể sử dụng kết quả nghiên cứu để ra quyết định và hoạch định chiến lược kinh doanh. Từ đó, cũng giúp nâng cao trải nghiệm khách hàng và tối đa hóa giá trị mang lại cho họ. Đồng thời, tăng cường lợi ích của doanh nghiệp và các bên có liên quan nói chung.

## Cấu trúc nghiên cứu

Đề tài được thực hiện theo cấu trúc 5 chương với nội dung chính như sau:

**Chương 1: Giới thiệu**

Đây là chương tổng quan và bao hàm những phần giới thiệu chung về đề tài như: lý do chọn đề tài, mục tiêu, đối tượng và phạm vi nghiên cứu, ý nghĩa khoa học và ý nghĩa thực tiễn của đề tài, cấu trúc bài nghiên cứu.

**Chương 2: Cơ sở lý thuyết**

Trong chương này, nhóm tác giả tổng hợp các công trình nghiên cứu trong và ngoài nước liên quan đến đề tài, cơ sở lý thuyết về phân tích cảm xúc, độ tin cậy của sản phẩm và trình bày về các phương pháp học máy được ứng dụng.

**Chương 3: Phương pháp nghiên cứu**

Đây là chương có nhiệm vụ quan trọng nhất với nội dung là mô tả bài toán trong nghiên cứu này, sau đó đề xuất giải pháp, và xây dựng mô hình giải pháp để giải quyết bài toán.

**Chương 4: Kết quả nghiên cứu và thảo luận**

Chương này trình bày kết quả thực nghiệm mà nhóm tác giả thu được, thêm vào đó là các phân tích về kết quả cũng như rút ra nhận xét..

**Chương 5: Kết luận và hướng phát triển**

Đây là chương tạm thời khép lại đề tài bằng cách đưa ra kết luận tổng quan nhất, tóm

tắt nội dung và kết quả của đề tài. Sau đó sẽ chỉ ra những ưu điểm, hạn chế và hướng

phát triển của đề tài trong tương lai.

# Cơ sở lý thuyết và các nghiên cứu liên quan

## Các nghiên cứu liên quan

### Nghiên cứu quốc tế

Từ bộ dữ liệu thuộc thu thập từ các sàn thương mại điện tử, Xiaoxin L. (2020) sử dụng 5 mô hình học máy khác nhau để phân tích và đưa ra các kết quả, so sánh một cách khách quan và đi đến kết luận rằng mô hình LightGBM cho ra kết quả cao nhất với độ chính xác là 98%, giá trị AUC là 96%.

Năm 2022, Priyanshi K. và các cộng sự đã ứng dụng các mô hình học máy và NLP để phân tích, tìm hiểu hành vi của khách hàng thông qua review và rating trên các sàn thương mại điện tử như Flipkart, Amazon,...với các sản phẩm thời trang nữ. Nhóm tác giá thực hiện phân tích với nhiều thông số và nhiều mô hình khác nhau để tìm hiểu liệu khách hàng có thích một sản phẩm nào đó hay không. Bên cạnh các mô hình học máy, các tác giả còn sử dụng thêm 2 thư viện của Python là Vader và Textblob để phân tích cụ thể hơn về cảm xúc trong một lượt đánh giá của khách hàng. Kết quả của bài nghiên cứu này cho thấy rằng, mô hình hồi quy Logistic (Logistic Regression) cho kết quả có độ chính xác cao nhất ở cả hai nhóm phân tích là bình luận (88,18%) và đánh giá (80,68%).

Cũng cùng ngành hàng thời trang nữ như trên, Manal L., Fayçal M. và Mohammed E. (2023) đã thực hiện phân tích bằng nhiều mô hình học máy có giám sát bao gồm k-Nearest Neighbors, Random Forest, Catboost, Logistic Regression. Sau phân tích, các tác giả đi đến kết luận các mô hình được đề xuất đều đạt độ chính xác từ 80% đến 90%, đặc biệt, mô hình hồi quy Logistic vượt trội hơn các mô hình khác về các tham số ma trận nhầm lẫn và cả điểm AUC.

Mặc dù các nghiên cứu trên sử dụng nhiều loại mô hình khác nhau nhưng nhìn chung đều chỉ ra rằng mô hình hồi quy Logistic có sự vượt trội hơn khi phân tích về đánh giá và bình luận của khách hàng trên các trang thương mại điện tử.

Tuy nhiên, trong nghiên cứu của Masfiq M. và các cộng sự (2023) khi phân tích dữ liệu được thu thập từ một nền tảng thương mại điện tử ở Bangladesh về lĩnh vực thời trang phụ nữ, nhóm tác giả sử dụng 6 mô hình bao gồm Random Forest, Support Vector Machine (SVM), Logistic Regression, Naive Bayes Classifier, Gradient Boosting, Long Short-Term Memory (LSTM). Một điểm thú vị trong bài nghiên cứu này chính là các tác giả đã đề cập đến việc lọc các đánh giá tiêu cực nhưng được xếp hạng 4 hoặc 5 sao và sau khi phân tích đi đến kết luận rằng mô hình Random Forest có độ chính xác cao nhất, lên đến 96%. Điều này cho thấy sự khác biệt với những nghiên cứu ở trên khi hầu hết các nghiên cứu trên đi đến kết luận mô hình hồi quy Logistic có độ chính xác cao hơn Random Forest.

### Nghiên cứu trong nước

Tác giả Hồ Trung Thành và các cộng sự (2021) đã sử dụng 2 mô hình học máy có giám sát là hồi quy Logistic và Random Forest để phân tích ý kiến khách hàng

trực tuyến với bộ dữ liệu từ Lazada. Mô hình được đề xuất có thể nhận biết được cảm xúc tích cực và tiêu cực trong ý kiến khách hàng đã ảnh hưởng 90% đến ứng dụng Lazada, điển hình tập trung vào các tính năng, chất lượng của ứng dụng.

Bài nghiên cứu của tác giả Đỗ Thành Đạt (2023) với đề tài “Ứng dụng phương pháp học máy vào phân tích cảm xúc khách hàng về dòng sản phẩm Herbalife” đã sử dụng 4 thuật toán bao gồm Naive Bayes, Logistic Regression, Support Vector Machine (SVM), XGBoost và kết quả là cả 4 mô hình đều có độ chính xác rất cao, từ 87%-89% trong đó Logistic Regression có độ chính xác cao nhất.

Tóm lại, hầu hết các nghiên cứu chỉ thực hiện được trên rất ít ngôn ngữ, nghiên cứu trong nước cũng chỉ xử lý ngôn ngữ tiếng anh, chưa mở rộng được thu thập và xử lý ngôn ngữ tiếng Việt. Bên cạnh đó, các nghiên cứu chỉ tập trung vào mô hình phân tích cảm xúc mang tính tích cực và tiêu cực, bỏ qua cảm xúc mang tính trung tính. Đây là một khuyết điểm lớn vì cảm xúc mang tính trung tính có ảnh hưởng lớn đến việc đánh giá độ tin cậy. Do đó nghiên cứu này sẽ cải thiện những hạn chế của các nghiên cứu trước, đánh giá toàn diện hơn trong nghiên cứu phân tích cảm xúc khách hàng.

## Cơ sở lý thuyết

### Phân tích cảm xúc

Phân tích cảm xúc, còn được gọi là khai thác thông tin, là một trong những lĩnh vực nghiên cứu đang hoạt động tích cực và phát triển nhanh chóng. Nó nhằm mục đích khai thác ý kiến ​​và thái độ của mọi người đối với các chủ đề, sự kiện, vấn đề, con người và sự vật khác nhau được truyền tải qua phương tiện truyền thông xã hội (chẳng hạn như blog, diễn đàn và trang mạng xã hội) thông qua đánh giá hoặc nhận xét văn bản. (Salukar & Pitale & Tikhe, 2023). Phân tích cảm xúc là một trong những cách tiếp cận thông thường đối với đánh giá của khách hàng, chủ yếu là để dự đoán thái độ của một đoạn văn bản là tích cực hay tiêu cực, ủng hộ hay phản đối (L. Zhang, Y. Zhou, X. Duan & R. Chen, 2018). Ngoài ra, việc phân tích ngữ nghĩa còn được sử dụng để phân tích đánh giá của khách hàng (Zhang, Y. Zhou, 2018; Bagheri, M. Saraee, 2013) cho các mục tiêu khác nhau như đo lường chất lượng dịch vụ thương mại điện tử (P. K. Sari, A. Alamsyah và S. Wibowo, 2018).

### API

1. **Khái niệm và cơ chế hoạt động của API**

* **Khái niệm:**

API là từ viết tắt của Application Programming Interface (Giao diện Lập trình Ứng dụng) là cơ chế cho phép 2 thành phần phần mềm giao tiếp với nhau bằng một tập hợp các định nghĩa và giao thức.

“Ứng dụng” đề cập đến mọi phần mềm có chức năng riêng biệt. “Giao diện” có thể được xem là một hợp đồng dịch vụ giữa 2 ứng dụng, xác định cách thức hai ứng dụng giao tiếp với nhau thông qua các yêu cầu và phản hồi.

* **Cơ chế hoạt động của API:**

Kiến trúc API thường được giải thích dưới dạng máy chủ và máy khách. Ứng dụng gửi yêu cầu được gọi là máy khách, còn ứng dụng gửi phản hồi được gọi là máy chủ.

API hoạt động theo 4 cách khác nhau, tùy vào thời điểm và lý do chúng được tạo ra:

|  |  |  |
| --- | --- | --- |
| **API** | **Cơ chế hoạt động** | **Đặc điểm** |
| API SOAP | Sử dụng Giao thức truy cập đối tượng đơn giản. Máy chủ và máy khách trao đổi thông điệp bằng XML. | Kém linh hoạt, trước đây được dùng phổ biến |
| API RPC | Máy khách hoàn thành một hàm (hoặc thủ tục) trên máy chủ còn máy chủ gửi kết quả về cho máy khách. | Lệnh gọi thủ tục từ xa |
| API Websocket | Sử dụng các đối tượng JSON để chuyển dữ liệu. API WebSocket hỗ trợ hoạt động giao tiếp hai chiều giữa ứng dụng máy khách và máy chủ. Máy chủ có thể gửi thông điệp gọi lại cho các máy khách được kết nối. | Hiệu quả cao |
| API REST | Máy khách gửi yêu cầu đến máy chủ dưới dạng dữ liệu. Máy chủ dùng dữ liệu đầu vào từ máy khách này để bắt đầu các hàm nội bộ và trả lại dữ liệu đầu ra cho máy khách. | Phổ biến và linh hoạt nhất trên web hiện nay. |

Bảng 2.1. Cơ chế hoạt động của API

1. **Web API**

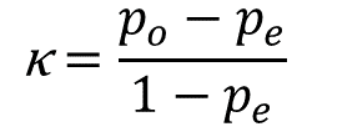
Web API một loại API được sử dụng cho các ứng dụng web, cho phép các ứng dụng khác truy cập và sử dụng các dịch vụ và tính năng của một trang web mà không cần truy cập vào giao diện người dùng của trang web đó. WebAPI thường được sử dụng để tạo ra các ứng dụng hoặc dịch vụ cho các trang web và các thiết bị di động.

### Fleiss’ Kappa

Fleiss’ Kappa là một phép đo thống kê được sử dụng để đánh giá độ tin cậy của thỏa thuận giữa một số người đánh giá (raters) khi gán xếp hạng phân loại cho một số mục hoặc phân loại mục. Đặc biệt, nó thường được sử dụng trong việc đo độ đồng thuận giữa nhiều nhà nghiên cứu khi phân loại các mẫu thành các nhóm hoặc phân loại.

Biến được sử dụng trong Fleiss Kappa là biến danh nghĩa (nominal variable)

Chỉ số Fleiss’ Kappa có thể được tính theo công thức sau:



Trong đó, là mức độ đồng thuận được quan sát của những người đánh giá

là mức độ đồng thuận mong đợi khi đánh giá ngẫu nhiên

Fleiss’ Kappa đo lường sự đồng nhất bằng cách so sánh sự đồng thuận giữa các nhà quan sát với sự đồng thuận ngẫu nhiên. Giá trị Kappa nằm trong khoảng từ -1 đến 1, với 1 biểu thị sự đồng thuận hoàn toàn, 0 là sự đồng thuận ngẫu nhiên, và giá trị âm thường chỉ ra sự không đồng thuận.

### Kỹ thuật nhúng từ - BERT Embedding

BERT (Bidirectional Encoder Representations from Transformers - Thể hiện Mã hóa Hai chiều từ Transformer) là một mô hình ngôn ngữ máy học được phát triển bởi Google, chủ yếu được sử dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP).

Kiến trúc của mô hình BERT là một kiến trúc đa tầng gồm nhiều lớp Bidirectional Transformer encoder.

BERT là một trong những tiến bộ lớn trong lĩnh vực NLP vì nó được thiết kế để đào tạo ra các vector đại diện cho ngôn ngữ văn bản thông qua ngữ cảnh 2 chiều. Trong khi các kiến trúc cũ thường chỉ xem xét từ trái sang phải hoặc ngược lại thì BERT chấp nhận ngữ cảnh toàn bộ câu đầu vào, giúp nó hiểu được ý nghĩa của từng từ dựa trên ngữ cảnh xung quanh chúng.

### K-Fold

K-Fold là một phương pháp cross-validation được sử dụng để đánh giá hiệu suất của mô hình máy học trên dữ liệu.

**Cross-validation**

Cross-validation là một phương pháp kiểm tra độ chính xác của 1 máy học dựa trên một tập dữ liệu học cho trước. Thay vì chỉ dùng một phần dữ liệu làm tập dữ liệu học thì cross-validation dùng toàn bộ dữ liệu để dạy cho máy.

3 phương pháp cross-validation phổ biến:

* **Hold-out:** phương pháp cross-validation đơn giản nhất. Trong phương pháp này, dữ liệu được chia một cách ngẫu nhiên thành hai phần: một tập dữ liệu học dùng để huấn luyện mô hình và một tập dữ liệu còn lại để kiểm tra.
* **K-Fold:** là phương pháp nâng cấp của Hold-out. Dữ liệu được chia thành k tập con (folds), quá trình học của máy có k lần và mô hình được huấn luyện trên k-1 folds và kiểm tra trên fold còn lại.
* **Leave-one-out:** tương tự như K-Fold nhưng tối đa hóa số tập con

### Máy học truyền thống

1. **Naive Bayes**

* **Lý thuyết Naive Bayes**

Naive Bayes còn được biết đến với những tên gọi khác như Simple Bayes, Independence Bayes hay phân loại Bayes. Phân loại Bayes được đánh giá cao khả năng mở rộng, đòi hỏi một số thông số tuyến tính trong số lượng các biến trong nhiều lĩnh vực khác nhau.

Một phân loại Naive Bayes dựa trên ý tưởng nó là một lớp được dự bằng các giá trị đặc trưng cho các thành viên của lớp đó. Các đối tượng là một nhóm trong các lớp nếu chúng có cùng các đặc trưng chung. Có thể có nhiều lớp rời rạc hoặc lớp nhị phân.

Các luật Bayes dựa trên xác suất để dự đoán chúng về các lớp có sẵn dựa trên các đặc trưng được trích rút. Trong phân loại Bayes, việc học được coi như xây dựng mô hình xác suất của các đặc trưng và sử dụng mô hình này để dự đoán phân loại cho một ví dụ mới.

Trường hợp đơn giản nhất là tạo ra các giả thuyết độc lập về đặc trưng đầu vào và độc lập có điều kiện với mỗi một lớp đã cho.

**Định lý Bayes:** Giả sử A và B là hai sự kiện đã xảy ra, xác suất điều kiện A khi biết trước điều kiện B được xác định bởi:

**P(A|B) = P(B|A).P(A)/P(B)**

Trong đó:

* P(A): Xác suất sự kiện A xảy ra.
* P(B): Xác suất sự kiện B xảy ra.
* P(B|A): Xác suất (có điều kiện) của sự kiện B xảy ra, nếu biết rằng sự kiện A đã xảy ra.
* P(A|B): Xác suất (có điều kiện) của sự kiện A xảy ra, nếu biết rằng sự kiện B đã xảy ra.

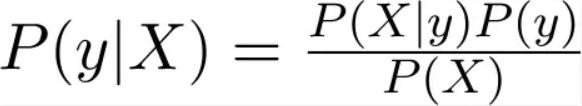
**Mô hình xác suất** cho phân loại là một mô hình điều kiện (C|F1,...,Fn)

**Xây dựng phân lớp từ mô hình xác suất**: Phân lớp Bayes kết hợp với luật quyết định tạo ra phân loại Naive Bayes. Bộ phân loại Naive Bayes là một hàm phân loại được định nghĩa:



* **Bộ phân loại Naive Bayes**

Naive Bayes là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học và nhiều lĩnh vực khác nhau. Thuật toán Naive Bayes dựa trên nguyên lý Bayes:



Ở đây, X là vector các đặc trưng, có thể viết dưới dạng X = (x1,x2,...,xn).

Trong mô hình Naive Bayes, có hai giả thiết được đặt ra:

* Các đặc trưng đưa vào mô hình là độc lập với nhau. Tức là sự thay đổi giá trị của một đặc trưng không ảnh hưởng đến các đặc trưng còn lại.
* Các đặc trưng đưa vào mô hình có ảnh hưởng ngang nhau đối với đầu ra mục tiêu.

Khi đó, kết quả mục tiêu y để P(y|X) đạt cực đại trở thành:



Chính vì hai giả thiết gần như không tồn tại trong thực tế trên, mô hình này mới được gọi là naive (ngây thơ). Tuy nhiên, chính sự đơn giản của nó với việc dự đoán rất nhanh kết quả đầu ra khiến nó được sử dụng rất nhiều trong thực tế trên những bộ dữ liệu lớn, đem lại kết quả khả quan. Một vài ứng dụng của Naive Bayes có thể kể đến như: lọc thư rác, phân loại văn bản, dự đoán sắc thái văn bản, ...

1. **Logistic Regression:**

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

Phương trình hồi quy logistic có thể được xây dựng với các giá trị 0 hoặc 1 được thu nhận từ kết quả đánh giá cảm quan của người tiêu dùng.

Trong đó, đầu vào là giá trị và đầu ra là F(x). Trong phân tích hàm nhiều biến, o+1x có thể được sửa đổi thành . Sau đó, khi được sử dụng trong các phương trình liên quan đến tỉ số odds với giá trị của các yếu tố dự báo, phương trình hồi quy tuyến tính sẽ trở thành hồi quy không tuyến tính với m biến, các thông số cho tất cả j = 0, 1, 2,.., m được ước tính.

1. **Support Vector Machine**

SVM là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên, thường thì SVM được sử dụng chủ yếu cho việc phân loại.

Support Vector Machine là bài toán đi tìm mặt phân cách sao cho *margin* tìm được là lớn nhất, đồng nghĩa với việc các điểm dữ liệu an toàn nhất so với mặt phân cách. Trong thuật toán này, chúng ta vẽ đồ thị dữ liệu là các điểm trong n chiều (với n là số lượng các tính năng) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.

Support Vectors hiểu một cách đơn giản là các đối tượng trên đồ thị tọa độ quan sát và Support Vector Machine là một biên giới để chia hai lớp tốt nhất.

### Deep Learning: CNN

**CNN** được viết tắt củaConvolutional Neural Network (mạng nơ-ron tích chập), là một trong những mô hình Deep Learning cực kỳ tiên tiến, cho phép xây dựng những hệ thống có độ chính xác cao và thông minh và được ứng dụng trong rất nhiều lĩnh vực khác nhau trong đời sống.

* **Cấu trúc của mạng CNN**

Mạng CNN là một tập hợp các lớp Convolution chồng lên nhau và sử dụng các hàm nonlinear activation như ReLU và tanh để kích hoạt các trọng số trong các node. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo. Trong mô hình mạng truyền ngược (feedforward neural network) thì mỗi neural đầu vào (input node) cho mỗi neural đầu ra trong các lớp tiếp theo. Đây được gọi là mạng kết nối đầy đủ (fully connected layer) hay mạng toàn vẹn (affine layer).

* 2 khía cạnh cần được quan tâm trong mô hình CNN là tính bất biến (Location Invariance) và tính kết hợp (Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các gốc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể.

### Mô hình Transformer

1. **PhoBert:**

PhoBERT là một mô hình BERT được huấn luyện dành riêng cho tiếng Việt và đạt được nhiều kết quả tốt cho nhiều nhiệm vụ trong xử lý ngôn ngữ này, cung cấp một thước đo cơ sở cho các bài toán về tiếng Việt. Do đó, đây là một mô hình huấn luyện đơn ngữ (monolingual language).

Có hai phiên bản của PhoBERT: PhoBERTbase (12 khối transformers) và PhoBERTlarge (24 khối transformers). Cả hai đều có dùng chung kiến trúc của BERTBase và BERTLarge. Nhóm tác giả huấn luyện mô hình trên 20GB dữ liệu tiếng Việt: 1GB các bài viết Wikipedia tiếng Việt và 19GB tin tức. Tuy nhiên, PhoBERT cần một bước tiền xử lý tách từ (ví dụ “tôi là sinh viên” → [“tôi”, “là”, “sinh\_viên”]) của một thư viện khác.

Qua thử nghiệm, mô hình PhoBERT cho kết quả tốt nhất hiện tại (SOTA), tốt hơn so với mô hình đa ngôn ngữ XML-R, khi thực hiện 4 nhiệm vụ NLP tiếng Việt bao gồm:

* POS (Part of Speech) - Gán nhãn từ loại;
* DP (Dependency Paser) - Phân tích sự phụ thuộc cú pháp;
* NER (Named Entity Recognition) - Nhận dạng thực thể;
* NLI (Natural Language Inference) - Suy luận ngôn ngữ tự nhiên.

1. **XLM-R**

XML-R là một mô hình huấn luyện đa ngôn ngữ được huấn luyện trong khoảng hơn 100 ngôn ngữ, trong đó có tiếng Việt.

Vì tiếng Việt có nhiều điểm không giống tiếng Anh nên các siêu tham số như epoch number, batch size (kích thước lô) hoặc tốc độ học máy sẽ khác nhau cho nên cần thử nghiệm nhiều lần để chọn ra thông số phù hợp nhất.

# Phương pháp nghiên cứu

## Mô tả bài toán

Sự phát triển mạnh mẽ của công nghệ đã giúp ngành thương mại điện tử trở thành một trong những lĩnh vực có tốc độ tăng trưởng nhanh nhất hiện nay. Khách hàng có xu hướng mua sắm trực tuyến nhiều hơn, và đánh giá sản phẩm là một yếu tố quan trọng ảnh hưởng đến quyết định mua hàng của người tiêu dùng. Tuy nhiên, trên đa số các sàn thương mại điện tử hiện nay, thông tin đánh giá sản phẩm đôi khi không đầy đủ, không thật sự chính xác hoặc bị ảnh hưởng bởi những yếu tố chủ quan khác. Điều này khiến người tiêu dùng gặp khó khăn trong việc dựa vào review đánh giá chất lượng sản phẩm để ra quyết định mua hàng.

Để giải quyết vấn đề trên, bài nghiên cứu này hướng đến việc xây dựng một mô hình dự đoán độ tin cậy của sản phẩm dựa vào phân tích cảm xúc bình luận của ngành hàng Thời trang trên sàn TMĐT Shopee.

* Input: Dữ liệu đầu vào của bài toán đến từ bình luận của người mua sản phẩm thuộc ngành hàng thời trang nữ trên sàn Shopee. Trong đó, cảm xúc của người mua được biểu diễn dưới 3 dạng: tích cực (positive), tiêu cực (negative) hoặc trung tính (neutral).
* Output: Kết quả đầu ra sẽ là rating sau khi huấn luyện với độ chính xác và tin cậy cao hơn.

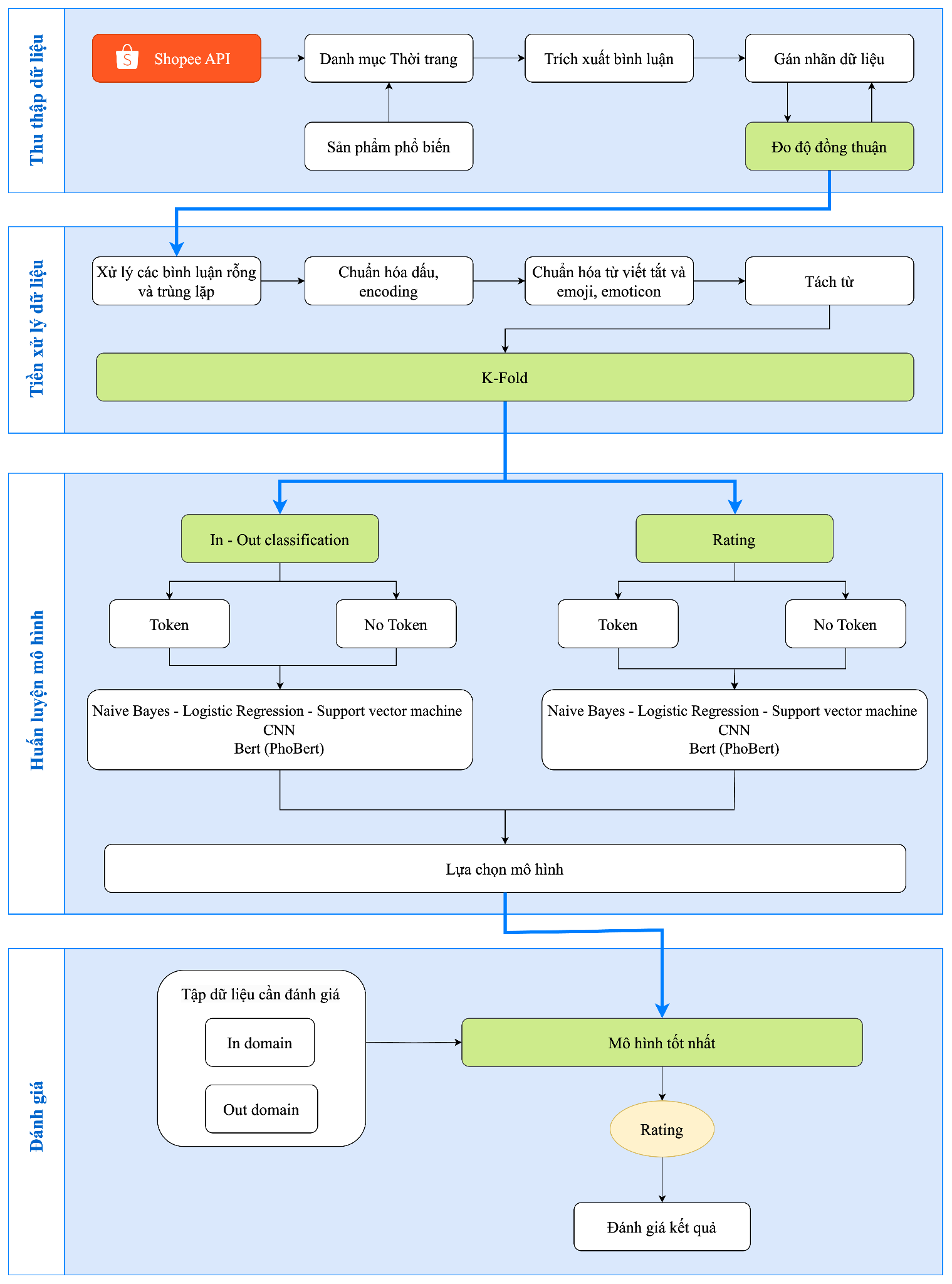
## Đề xuất giải pháp

### Giải pháp

Để giải quyết vấn đề cũng như thực hiện đề tài “Dự đoán độ tin cậy của sản phẩm dựa vào phân tích cảm xúc bình luận của ngành hàng Thời trang trên sàn Thương mại điện tử Shopee”, nhóm tiến hành phân tích thực trạng ngành thời trang ở thị trường Việt Nam và đo lường cảm xúc của người dùng thông qua các bình luận về sản phẩm. Liệu đánh giá của người dùng đối với sản phẩm đó có thật sự đáng tin cậy hay không và người dùng có thể dựa vào đánh giá trung bình của sản phẩm đó để đưa ra quyết định tiêu dùng được không? Cụ thể giải pháp được thực hiện như sau:

* Thực hiện thu thập dữ liệu bình luận và đánh giá của các sản phẩm thời trang trên sàn thương mại điện tử Shopee.
* Tập hợp dữ liệu và tiến hành tiền xử lý đối với những dữ liệu bình luận của sản phẩm.
* Đặt ra các quy tắc và tiến ành gán nhãn dữ liệu theo 3 nhãn (Tích cực, Tiêu cực và Trung tính).
* Đo lường độ đồng thuận giữa những người thực hiện gán nhãn.
* Tiến hành phân tích, thống kê và đưa ra cái nhìn tổng quát nhất đối với bộ dữ liệu.
* Xử lý nhúng từ từ bộ dữ liệu bình luận.
* Thực hiện huấn luyện mô hình và chọn mô hình tốt nhất.
* Ứng dụng mô hình vào bộ dữ liệu để xác định độ tin cậy của sản phẩm.
* Kiểm tra mô hình với bộ dữ liệu thuộc lĩnh vực khác để đánh giá tính đa nhiệm của mô hình.
* Tổng hợp kết quả và đánh giá.

### Xây dựng mô hình giải pháp



Hình 3.1. Mô hình giải pháp

# Kết quả nghiên cứu và thảo luận

## Mô tả cách lấy dữ liệu

Với mục tiêu phân tích cảm xúc bình luận của ngành hàng Thời trang trên sàn TMĐT Shopee, dữ liệu được thu thập dựa vào các bình luận, đánh giá trên sàn thương mại điện tử Shopee.vn. Dữ liệu này là công khai và tập trung vào danh mục thời trang.

Do sàn thương mại điện tử Shopee.vn có áp dụng nhiều cơ chế để bảo mật thông tin, vì vậy các thư viện Python thường dùng để truy cập và trích xuất dữ liệu (VD: thư viện Beautifulsoup, Web Scraper, v.v.) sẽ gặp khó khăn khi lấy dữ liệu từ sàn thương mại điện tử này. Do đó, để thu thập dữ liệu, nhóm đã sử dụng API (tính đến ngày nhóm thực hiện nghiên cứu là tháng 2 năm 2024) do Shopee cung cấp qua các bước:

**Bước 1**: Import thư viện requests. Requests là một thư viện bên ngoài được sử dụng để gửi yêu cầu HTTP và nhận phản hồi từ các trang web. Sử dụng thư viện requests để gửi HTTP request và nhận dữ liệu JSON từ API Shopee.



Hình 4.1: Import thư viện requests

**Bước 2**: Xác định những thông tin cơ bản của sản phẩm cần trích xuất: Tạo list selection\_prop bao gồm các thông tin:

* itemid: id của sản phẩm
* shopid: id của shop
* name: tên sản phẩm
* currency: đơn vị tiền tệ
* price: giá sản phẩm
* price\_before\_discount: giá gốc sản phẩm
* raw\_discount: giảm giá
* item\_rating: số điểm trung bình đánh giá của sản phẩm



Hình 4.1. Xác định thông tin cơ bản của sản phẩm cần trích xuất

**Bước 3**: Tại mỗi sản phẩm, tạo list selection\_cmt\_prop bao gồm thông tin về bình luận, đánh giá của sản phẩm đó

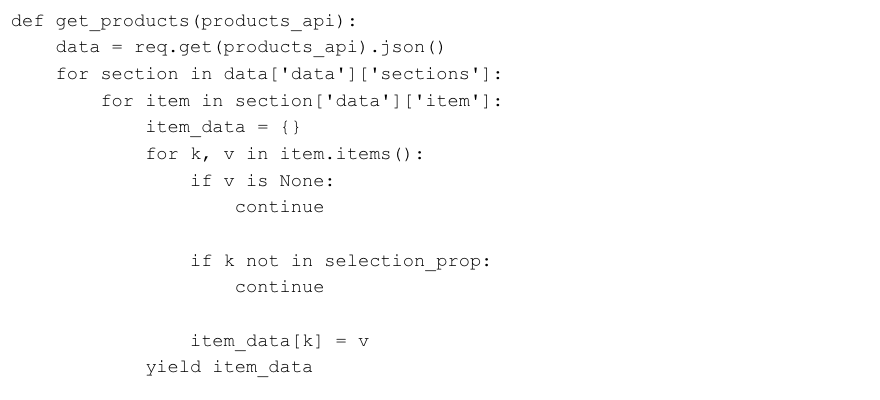
* cmtid: id của bình luận đánh giá sản phẩm
* comment: đoạn văn bản đánh giá sản phẩm
* rating\_star: số điểm mỗi khách hàng đánh giá



Hình 4.2. Xác định thông tin về bình luận và đánh giá

**Bước 4**: Hàm get\_products(products\_api)

* products\_api: URL API để lấy thông tin về các sản phẩm từ Shopee.
* Gửi một HTTP request đến products\_api để nhận dữ liệu JSON chứa thông tin về các sản phẩm Shopee.
* Sử dụng vòng lặp để trích xuất thông tin của từng sản phẩm từ dữ liệu nhận được.
* Sử dụng một bộ lọc (selection\_prop) để chỉ giữ lại các trường thông tin cần thiết đã liệt kê ở bước 2.
* Sử dụng generator (yield) để trả về từng sản phẩm một, giúp tiết kiệm bộ nhớ khi lấy dữ liệu hàng loạt.



Hình 4.3. Lấy API sản phẩm

**Bước 5**: Hàm get\_comments

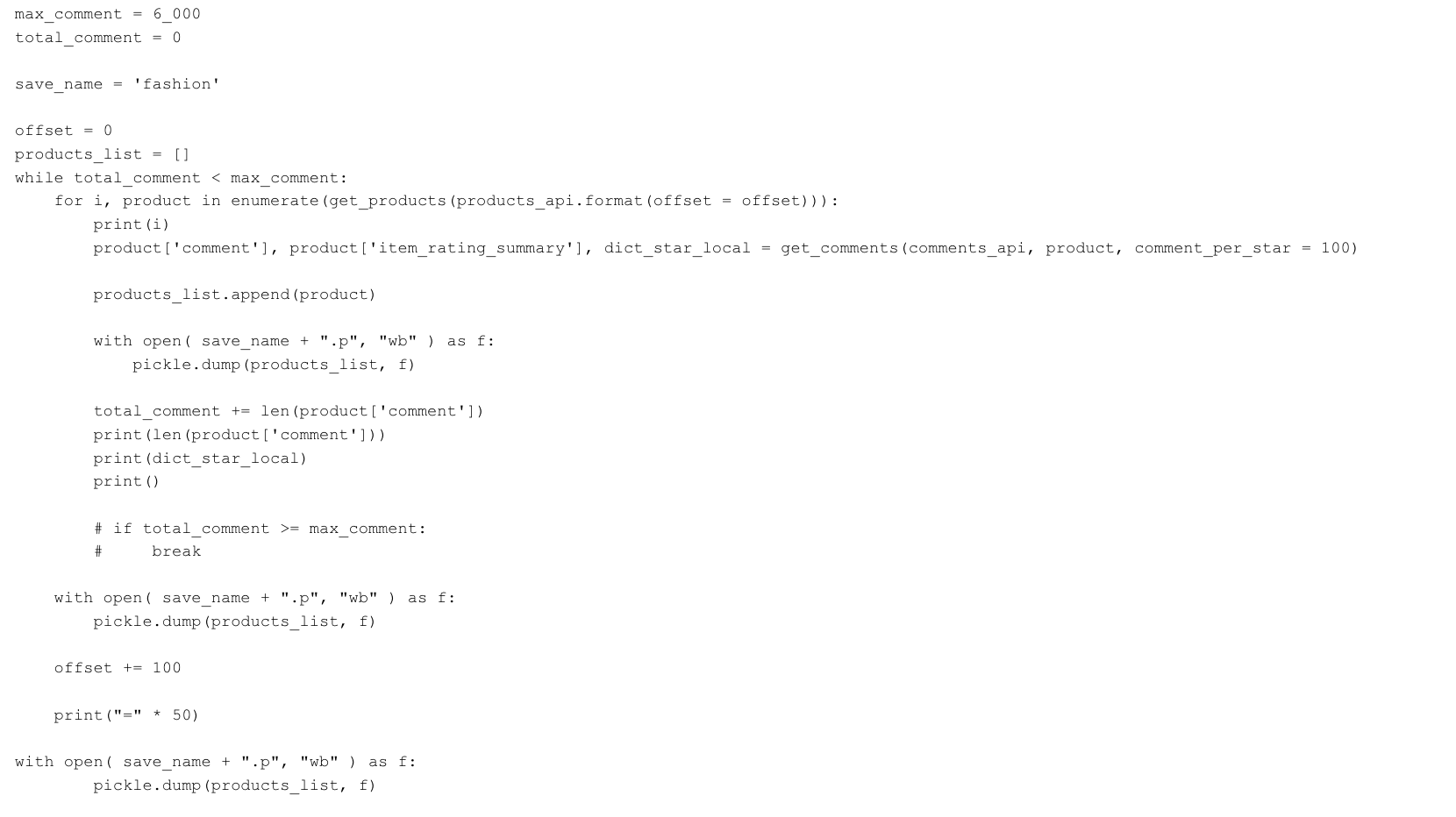
* comments\_api: URL API để lấy thông tin về đánh giá và bình luận của sản phẩm từ Shopee.
* product: Thông tin về sản phẩm được trả về từ hàm get\_products.
* comment\_per\_star: Số lượng bình luận muốn lấy cho mỗi sao. Ở nghiên cứu này, nhóm lấy 100 bình luận / mỗi sao.
* Gửi các yêu cầu API để lấy đánh giá và bình luận của sản phẩm theo từng sao.
* Sử dụng vòng lặp để lấy dữ liệu theo từng trang (mỗi trang có 50 bình luận).
* Trả về danh sách bình luận, tổng kết đánh giá sản phẩm và số lượng đánh giá theo từng sao.



Hình 4.4. Lấy API đánh giá và bình luận

**Bước 6**: Vòng lặp while

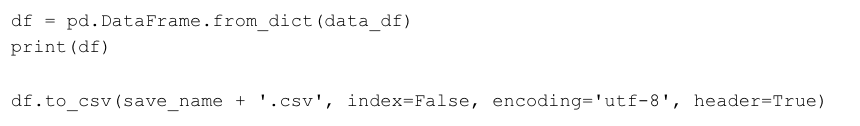
* Vòng lặp while này được sử dụng để lấy dữ liệu từ API Shopee, đánh giá và bình luận của các sản phẩm, và lưu trữ thông tin vào danh sách sản phẩm products\_list.
* Vòng lặp sẽ tiếp tục cho đến khi tổng số lượng bình luận total\_comment đạt đến giới hạn được đặt max\_comment = 100.
* offset được sử dụng để theo dõi vị trí trang hiện tại của danh sách sản phẩm trong kết quả trả về từ API.
* Mỗi lần lặp, get\_products được gọi để lấy thông tin về các sản phẩm từ trang hiện tại.
* Mỗi sản phẩm sau đó được chuyển qua hàm get\_comments để lấy thông tin đánh giá và bình luận.
* Thông tin được thêm vào danh sách products\_list, và danh sách này được lưu vào tệp pickle sau mỗi lần lặp để tránh mất dữ liệu nếu quá trình thực thi bị gián đoạn.
* total\_comment được cập nhật với số lượng bình luận mới để kiểm tra điều kiện dừng vòng lặp.
* Vòng lặp cũng cập nhật giá trị offset để chuyển đến trang tiếp theo của danh sách sản phẩm.



Hình 4.5. Vòng lặp while lấy tối đa 100 đánh giá và bình luận với mỗi sao

**Bước 7**: Chuyển đổi dữ liệu

* Tập tin data\_df được tạo để lưu trữ tất cả dữ liệu liên quan đến bình luận và đánh giá sản phẩm.
* Chuyển đổi tất cả dữ liệu bình luận và đánh giá bằng tiếng Việt thành một tệp CSV với tên save\_name.



Hình 4.6. Lưu dữ liệu thành tệp CSV

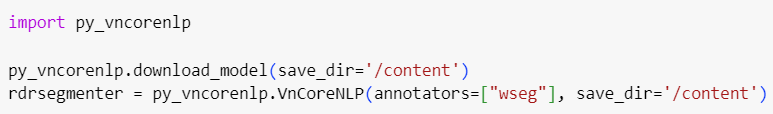
## Tiền xử lý dữ liệu

Bộ dữ liệu gốc sau khi lấy về luôn sẽ có những thành phần gây nhiễu quá trình phân tích cũng như là ảnh hưởng đến độ chính xác của kết quả đưa ra, vì vậy tiền xử lý dữ liệu là quy trình cần thiết để làm nền tảng thực hiện cho các bước sau này.

* **Import các thư viện sử dụng trong quá trình tiền xử lý**

****

Hình 4.7. Import thư viện emot

****

Hình 4.8. Import thư viện py\_vncorenlp

****

****

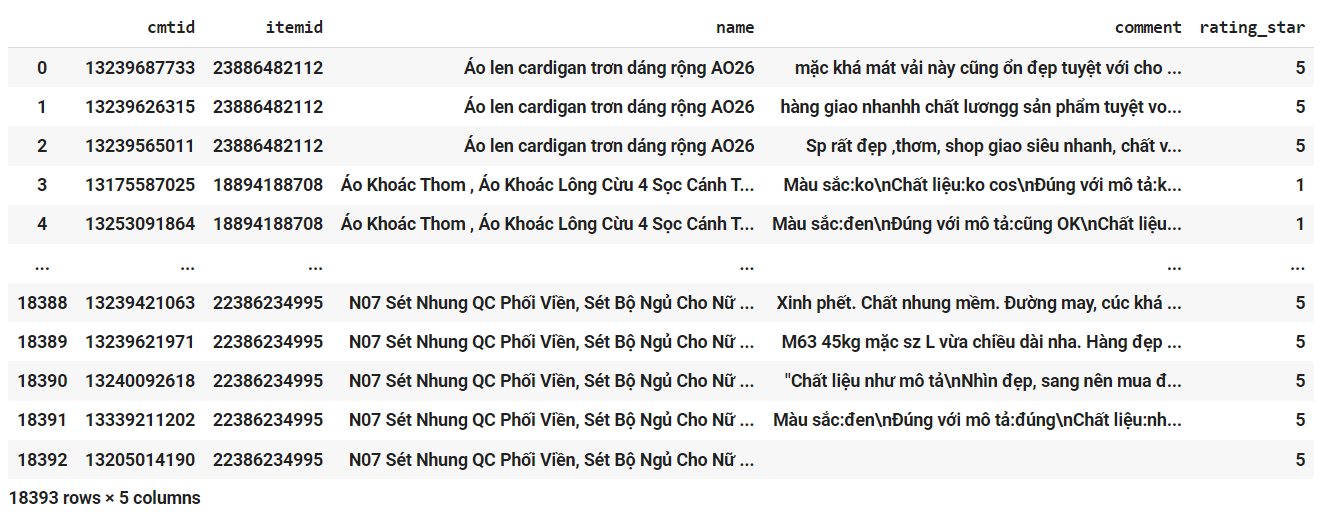
Hình 4.9. Import thư viện pickle

Trong đó bao gồm:

* Pandas - thư viện được sử dụng để xử lý và phân tích dữ liệu
* emot - hỗ trợ việc xử lý cảm xúc trong văn bản bằng cách nhận diện và thay thế các biểu cảm cảm xúc (emoji) trong văn bản thành các biểu tượng chữ thường
* py\_vncorenlp - là một thư viện cung cấp các chức năng như phân đoạn từ, gán nhãn từ loại, và nhận dạng thực thể trong văn bản tiếng Việt
* pickle - là một cách để lưu trữ và truy xuất các đối tượng Python dưới dạng biểu diễn nhị phân
* regex - là thư viện hỗ trợ việc thực hiện các thao tác xử lý chuỗi sử dụng biểu thức chính quy (regular expressions)
* underthesea - là thư viện xử lý ngôn ngữ tự nhiên cho tiếng Việt
* ftfy - Ftfy (fix text for you) là một thư viện giúp sửa các vấn đề liên quan đến mã hóa và giải mã văn bản, đảm bảo rằng văn bản được hiển thị một cách chính xác
* unicodedata - cung cấp các hàm để thực hiện các phép toán trên ký tự Unicode
* **Chuyển đổi dữ liệu lồng nhau thành một DataFrame của Pandas**



Hình 4.10. Chuyển đổi dữ liệu thành DataFrame

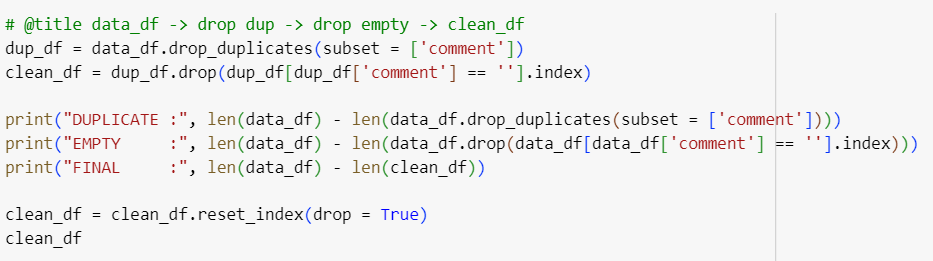


Hình 4.11. Dữ liệu gốc thu được sau khi lấy dữ liệu

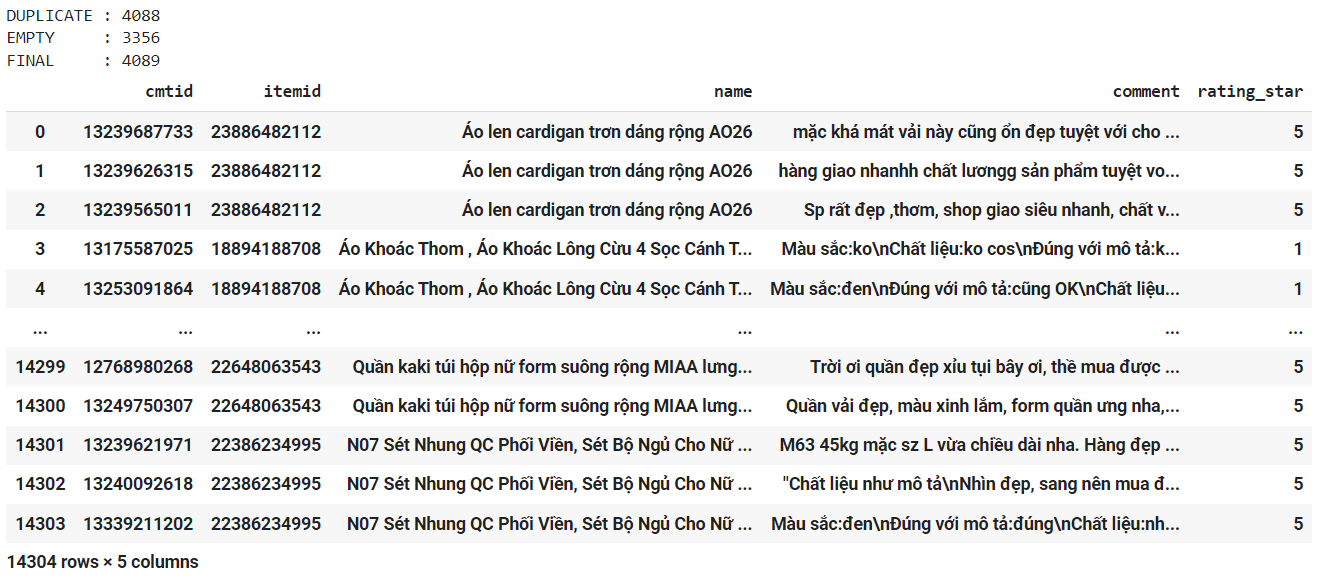
Mỗi dòng trong bộ dữ liệu này liên kết với một sản phẩm cụ thể. Cột itemid chứa mã sản phẩm mà dòng đánh giá liên quan đến. Cột name chứa tên của sản phẩm đó. Đồng thời, mỗi dòng trong bộ dữ liệu này cũng chứa một đánh giá cụ thể về sản phẩm nằm trong cột comment.

Trong đó:

* cmtid: Mã đánh giá là một số duy nhất được Shopee gán cho mỗi đánh giá. Mã này được sử dụng để tham chiếu đến đánh giá trong các hệ thống của Shopee.
* itemid: Mã sản phẩm là một số duy nhất được Shopee gán cho mỗi sản phẩm. Mã này được sử dụng để tham chiếu đến sản phẩm trong các hệ thống của Shopee.
* name: Tên sản phẩm là tên của sản phẩm được đánh giá. Tên này được cung cấp bởi người bán sản phẩm.
* comment: Nội dung đánh giá là văn bản được người mua viết để đánh giá sản phẩm. Nội dung đánh giá có thể bao gồm các thông tin như chất lượng sản phẩm, dịch vụ khách hàng, thời gian giao hàng,...
* rating\_star: Mức độ đánh giá chất lượng sản phẩm của người mua, được đánh giá từ 1 sao đến 5 sao.
* **Thực hiện loại bỏ các bình luận bị rỗng hoặc bị lặp lại**



Hình 4.12. Thực hiện loại bỏ các bình luận bị rỗng hoặc lặp lại



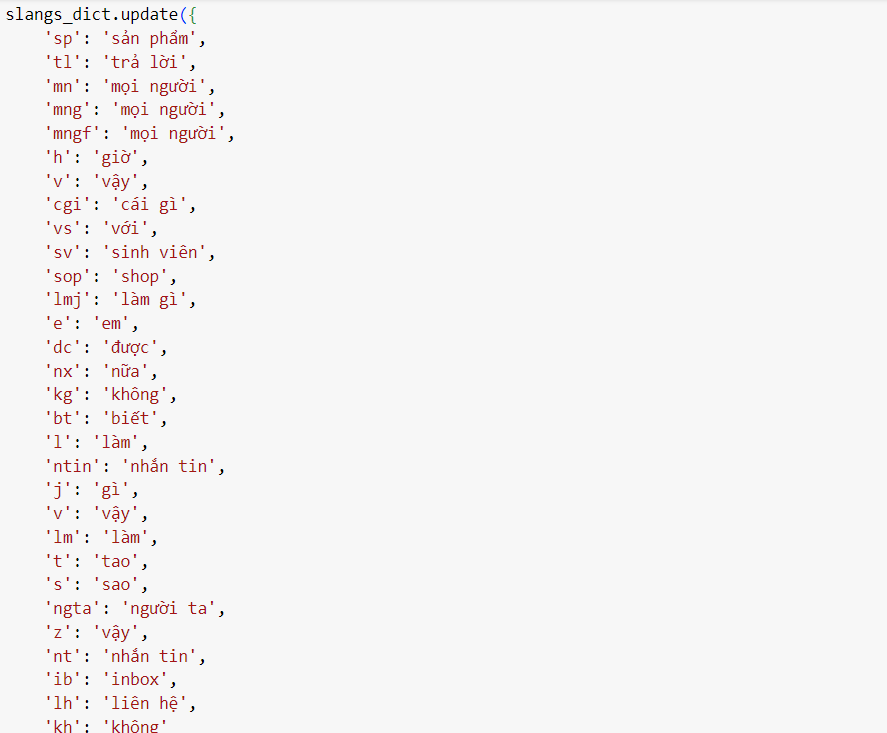
Hình 4.13. Bộ dữ liệu sau khi đã loại bỏ bình luận bị rỗng và lặp lại

Các bình luận rỗng hoặc lặp lại sẽ là một yếu tố gây ảnh hưởng không hề nhỏ đến kết quả phân tích, đồng thời khiến quá trình phân tích trở nên khó khăn, vì vậy loại bỏ chúng là bước quan trọng trong quá trình tiền xử lý dữ liệu. Sau khi loại bỏ, như kết quả trên, các dòng bị loại bỏ vì bị trùng lặp và bị rỗng là 4088 dòng, do đó bộ dữ liệu còn lại 14304 dòng.

* **Tạo từ điển (slangs\_dict) ánh xạ các từ ngữ viết tắt hoặc từ lóng (teencode)**



Hình 4.14. Tạo từ điển ánh xạ các từ ngữ viết tắt hoặc từ lóng (teencode)



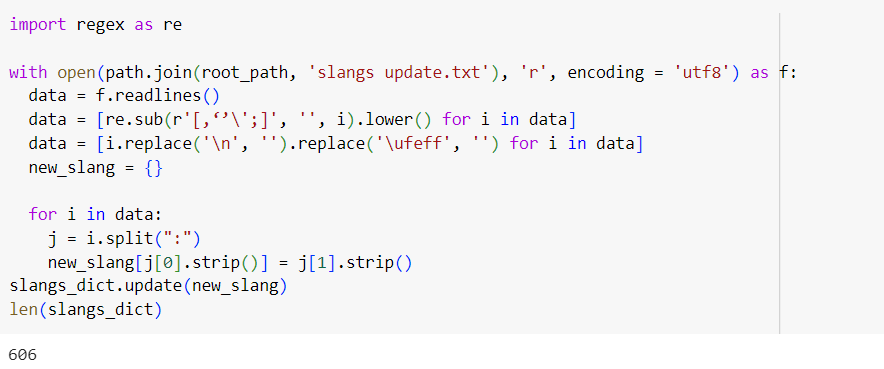
Hình 4.15. Cập nhật các từ viết tắt, từ lóng vào trong từ điển đã tạo



Hình 4.16. Kết quả số lượng từ viết tắt và từ lóng (teencode) trong từ điển

Các bình luận trên Shopee thường chứa các từ viết tắt hoặc từ lóng (teencode), vì vậy nhóm tạo từ điển này để chuyển thành từ ngữ chuẩn để thuận lợi hơn trong quá trình phân tích. Sau khi thực hiện nhóm thu được một từ điển từ ngữ viết tắt và teencode có tổng cộng là 521 từ.

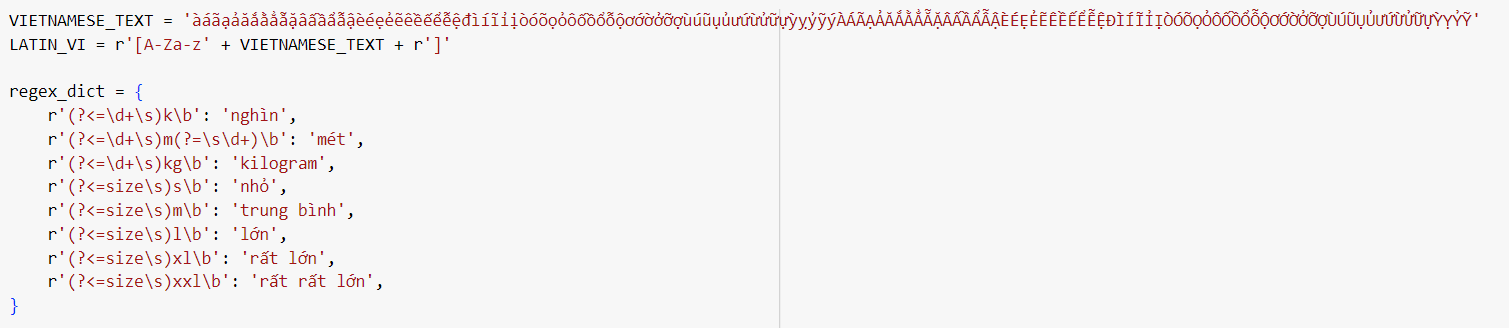
* **Cập nhật từ điển từ một tệp văn bản khác**



Hình 4.17. Cập nhật từ điển từ tệp văn bản khác

Nhóm tiếp tục cập nhật vào từ điển slangs\_dict đã tạo bằng cách đọc dữ liệu từ một tệp tin văn bản 'slangs update.txt' và thêm vào các ánh xạ mới. Từ điển về từ ngữ viết tắt và teencode cuối cùng của nhóm có 606 từ.

* **Xây dựng từ điển về kích thước của mặt hàng thời trang**



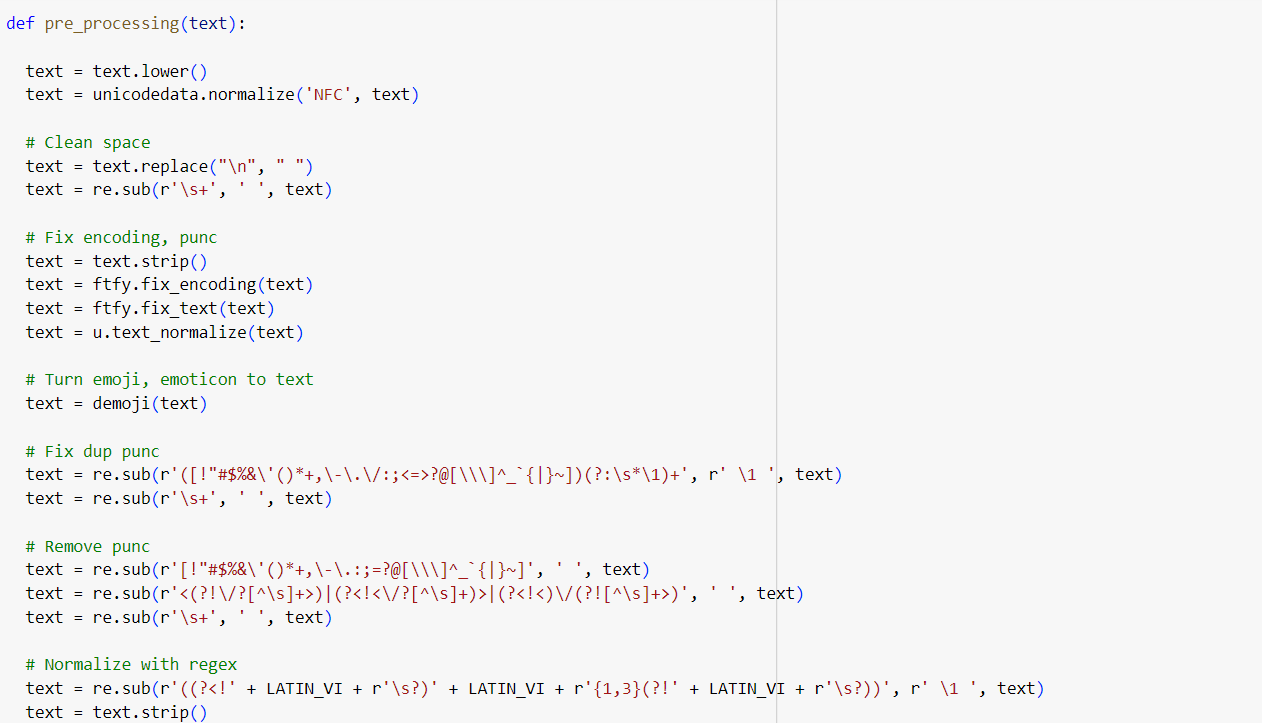
Hình 4.18. Xây dựng từ điển về kích thước của mặt hàng thời trang

Do tính chất đặc biệt của ngành hàng Thời trang, các bình luận có thể chứa thông tin về kích thước, chiều cao, cân nặng của khách hàng dưới dạng viết tắt như S, M, L. Đế trách việc chuyển đổi từ viết tắt bị sai, nhóm tiếng hành xây dựng từ điển cho các nhóm từ này.

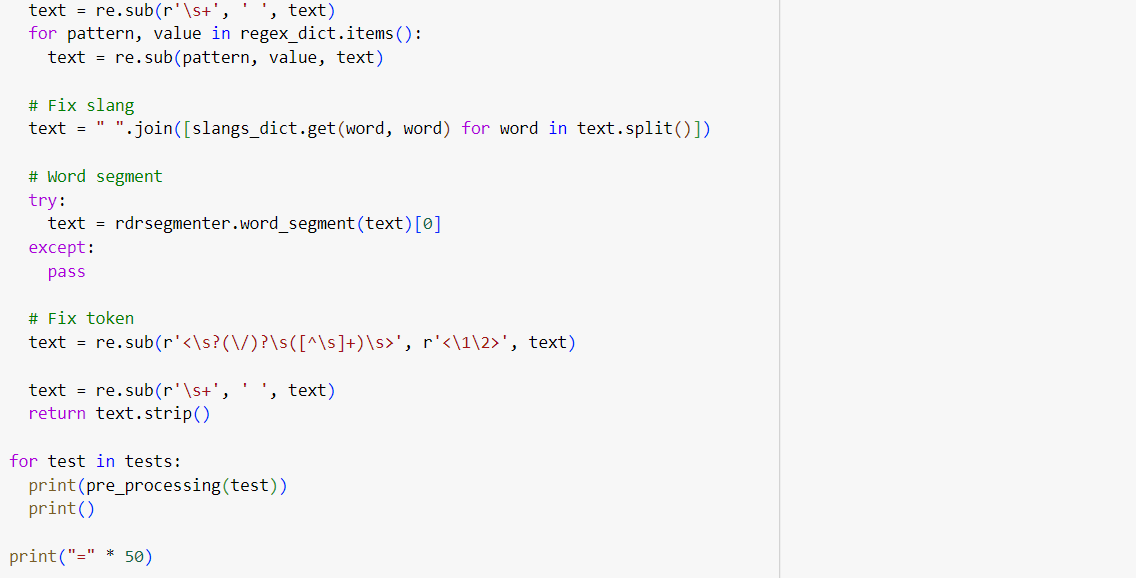
* **Chuẩn hoá và làm sạch dữ liệu văn bản của bình luận**



Hình 4.19. Chuyển đổi emoji và emoticon thành văn bản



Hình 4.20. Chuẩn hoá dữ liệu văn bản của bình luận (1)

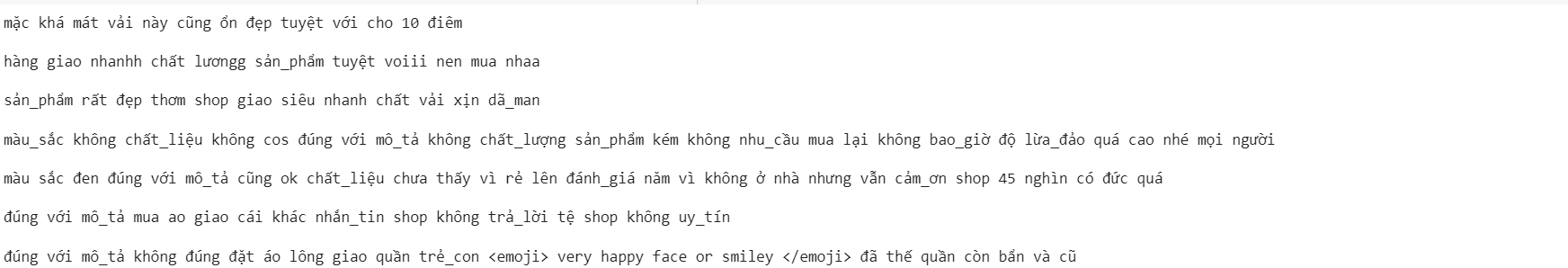


Hình 4.21. Chuẩn hoá dữ liệu văn bản của bình luận (2)

Tiếp tục với quá trình tiền xử lý dữ liệu, nhóm sử dụng hàm demoji để chuyển đổi emoji và emoticon thành văn bản. Và sử dụng hàm pre\_processing để tiền xử lý nội dung của comment bao gồm các công đoạn như sau:

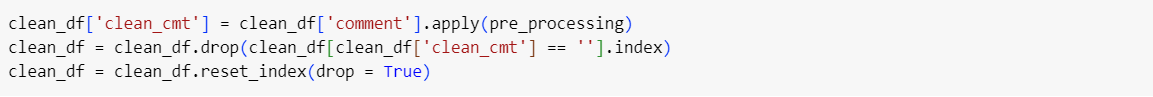
* Chuyển đổi toàn bộ văn bản thành chữ thường
* Chuẩn hóa unicode để đảm bảo rằng các ký tự được biểu diễn theo một cách duy nhất
* Thay thế ký tự xuống dòng bằng khoảng trắng
* Loại bỏ khoảng trắng thừa
* Sửa các lỗi chính tả và biểu hiện văn bản
* Chuẩn hóa văn bản sử dụng underthesea
* Loại bỏ dấu câu và các ký tự đặc biệt khác
* Sửa các từ viết tắt bằng cách thay thế chúng theo từ điển slangs\_dict
* Sử dụng py\_vncorenlp để tách từ

Cuối cùng hàm sẽ trả về văn bản đã được xử lý như sau:

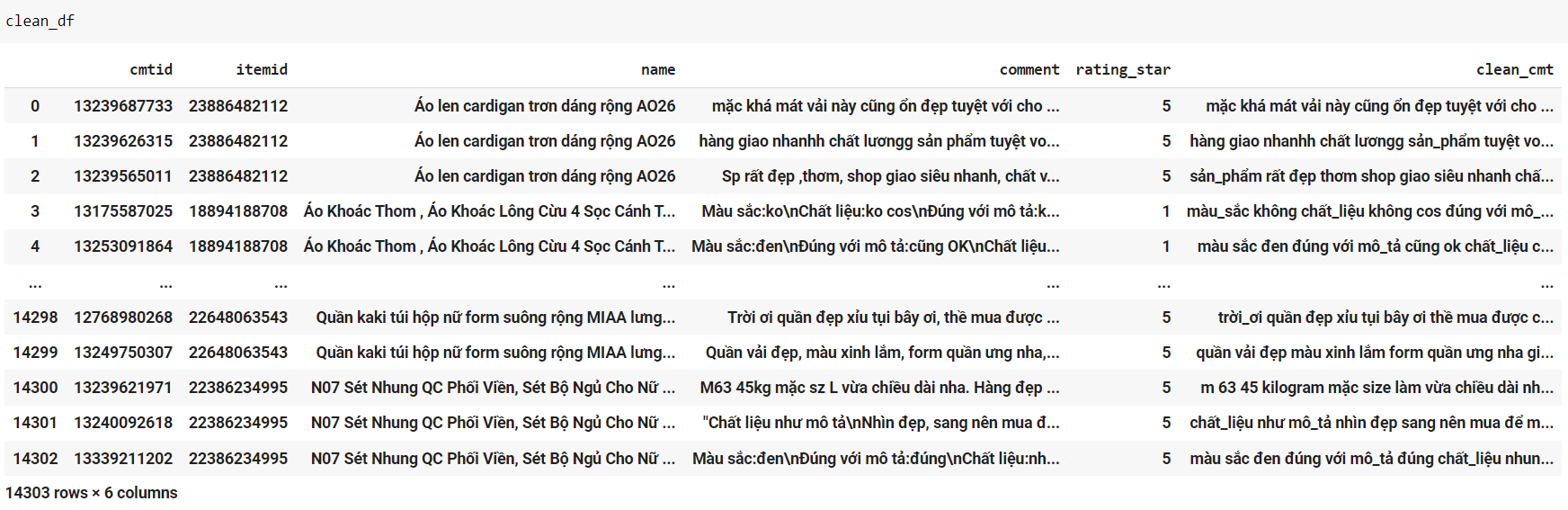


Hình 4.22. Dữ liệu bình luận sau khi đã được chuẩn hoá và làm sạch

* **Thêm bình luận đã xử lý và làm sạch vào DataFrame**



Hình 4.23. Thêm bình luận đã được xử lý và làm sạch vào DataFrame



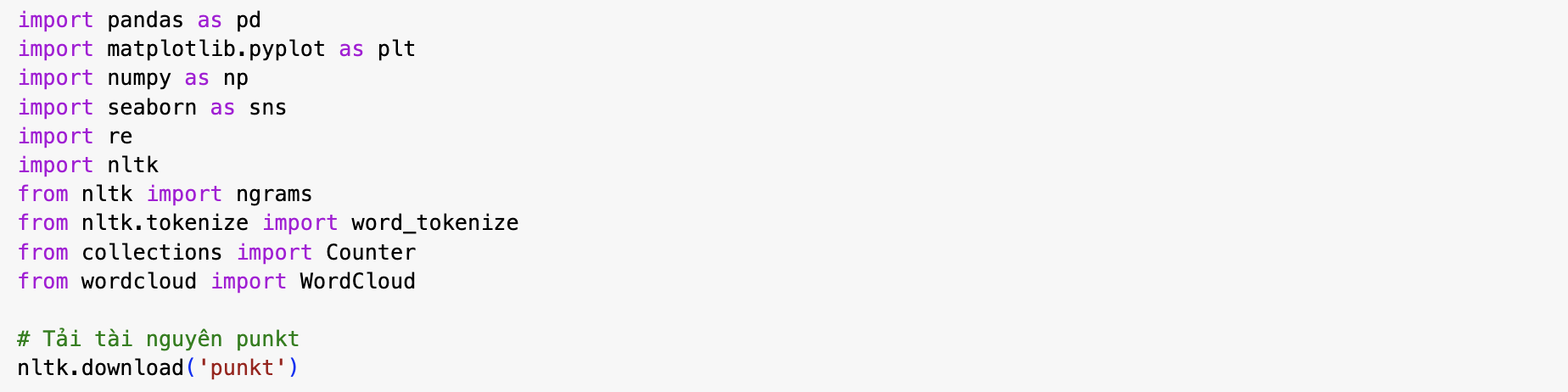
Hình 4.24. Bộ dữ liệu thu được sau khi đã thực hiện quá trình tiền xử lý

Sau khi thực hiện xong quá trình tiền xử lý dữ liệu, nhóm thu được thêm 1 cột dữ liệu 'clean\_cmt' - dữ liệu bình luận đã được chuẩn hoá, xử lý và làm sạch vào bộ dữ liệu để có thể tiếp tục tiến hành các bước phân tích tiếp theo.

## EDA

Với đề tài **“Đề xuất phân loại và điểm đánh giá cho bình luận của ngành hàng thời trang trên sàn thương mại điện tử Shopee”,** nhóm sẽ tập trung phân tích 2 cột dữ liệu là ‘rating\_star’ và phần bình luận đã được xử lý là ‘clean\_cmt'. Do đó, ở phần này nhóm tập trung khai phá dữ liệu của 2 cột dữ liệu trên để có cái nhìn tổng quan về chúng cũng như mối liên hệ giữa 2 cột giữ liệu này.

* **Import các thư viện để sử dụng cho quá trình EDA:**

****

Hình 4.25. Thư viện sử dụng trong quá trình EDA

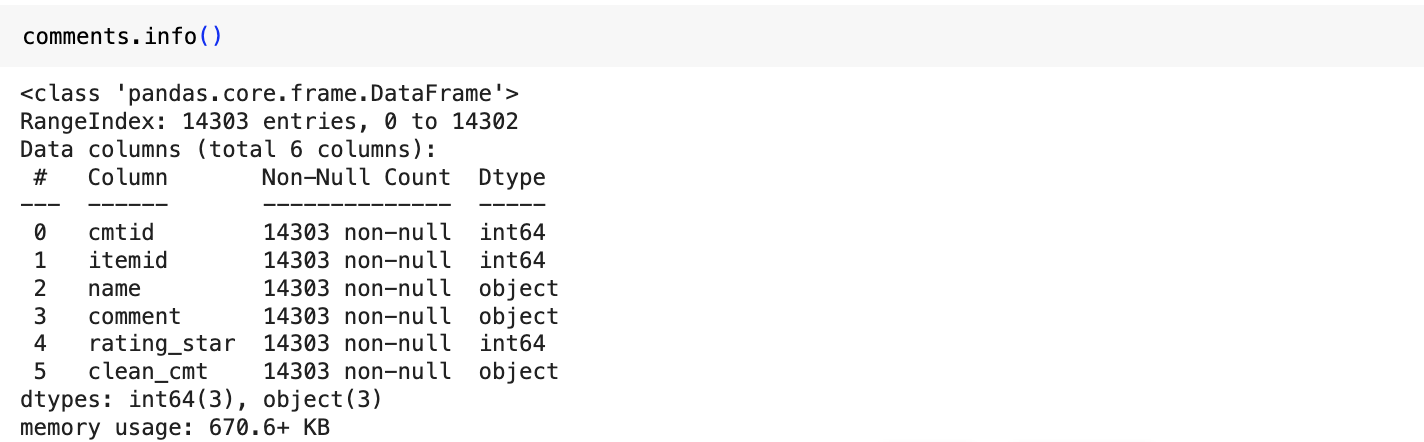
Các thư viện bao gồm:

* Pandas: pandas là một thư viện mạnh mẽ cho xử lý và phân tích dữ liệu, cung cấp các cấu trúc dữ liệu như DataFrame, giúp thực hiện các thao tác như lọc, nhóm, và tính toán trên dữ liệu.
* Matplotlib: matplotlib là một thư viện trực quan hóa dữ liệu, cho phép tạo các biểu đồ và đồ thị. Pyplot là một module trong matplotlib giúp tạo các biểu đồ dễ dàng.
* NumPy: numpy là một thư viện chuyên về tính toán số học trên mảng và ma trận. Nó cung cấp các hàm và công cụ linh hoạt giúp thực hiện các phép toán số học một cách hiệu quả.
* NLTK: nltk (Natural Language Toolkit) là một thư viện phổ biến cho xử lý ngôn ngữ tự nhiên. Trong trường hợp này nhóm sử dụng nltk để tải tài nguyên punkt, một tập hợp các mô hình được sử dụng cho việc tách từ và phân đoạn câu
* Seaborn: seaborn là một thư viện trực quan hóa dữ liệu xây dựng trên matplotlib. Nó cung cấp các chủ đề và hàm trực quan hóa cao cấp hơn, giúp làm cho biểu đồ trở nên đẹp mắt và dễ đọc hơn.
* WordCloud: wordcloud là một thư viện để tạo hình ảnh word cloud, trong đó kích thước của từ thể hiện tần suất xuất hiện của từ đó trong văn bản.
* **Đọc dữ liệu từ bộ dữ liệu đã được tiền xử lý**

****

Hình 4.26. Dữ liệu được sử dụng để khai phá dữ liệu

* **Kiểm tra thông tin bộ dữ liệu**

****

Hình 4.27. Thông tin bộ dữ liệu

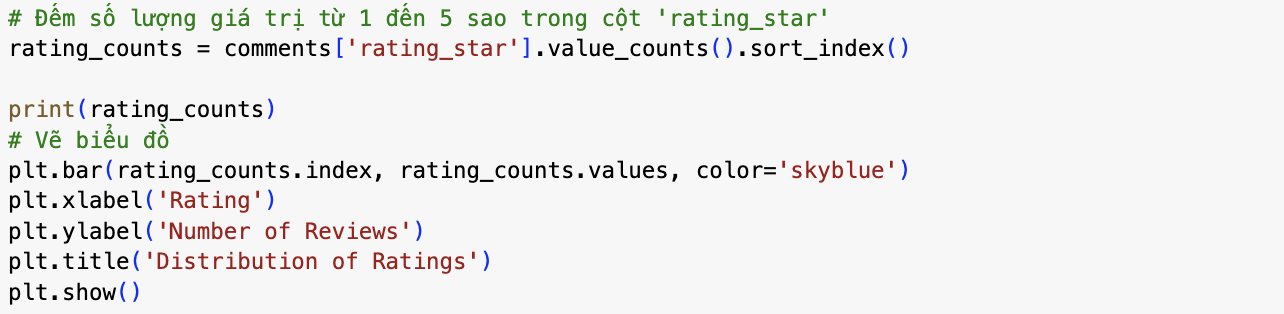
Bộ dữ liệu được sử dụng EDA là bộ dữ liệu đã được tiền xử lý, do vậy không có dòng dữ liệu nào bị rỗng. Bộ dữ liệu bao gồm 14303 dòng và 5 cột.

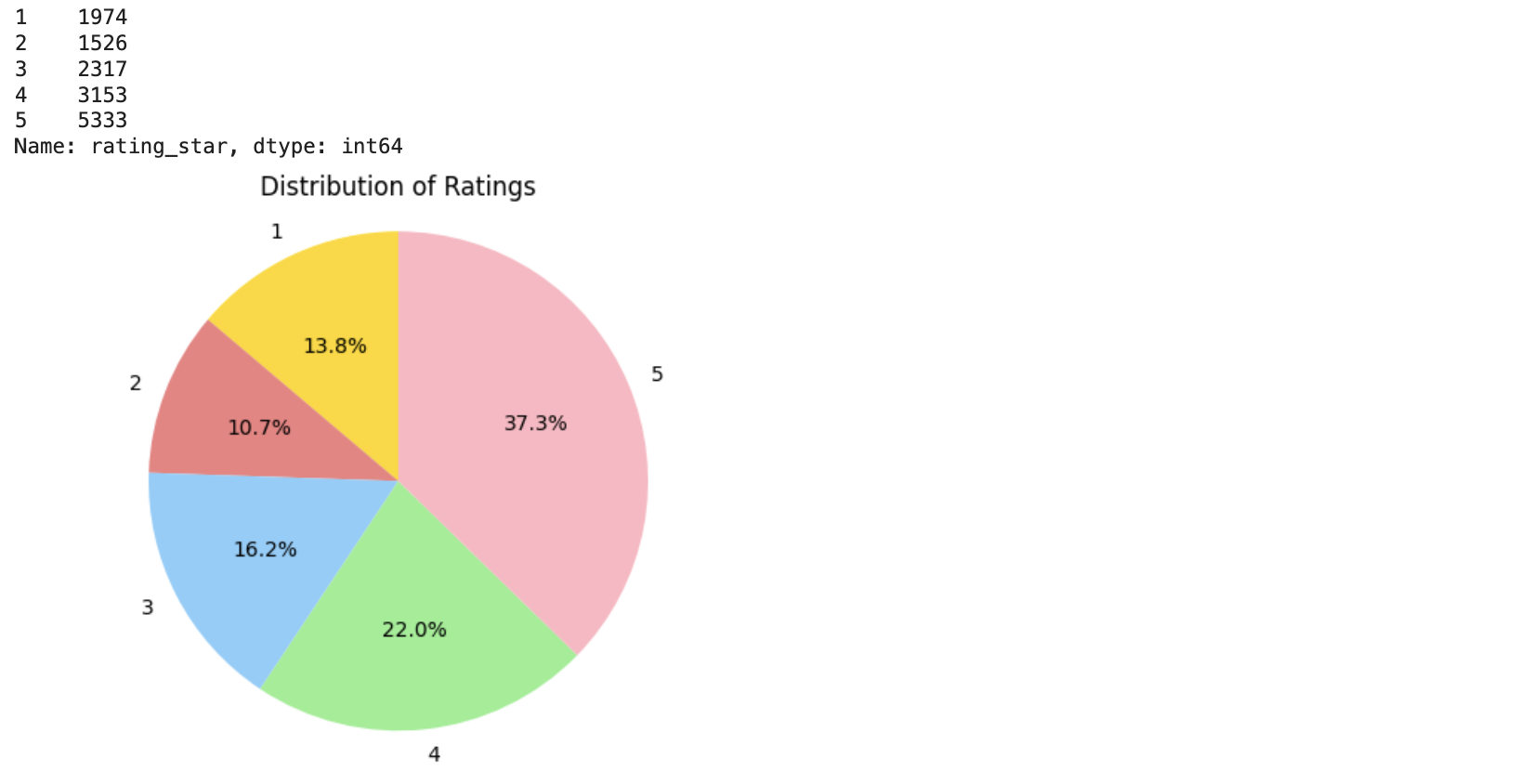
* **Mô tả cột dữ liệu ‘rating\_star’**



Hình 4.28. Mô tả cột dữ liệu rating star

Dữ liệu trong cột ‘rating\_star' cho thấy với 14303 đánh giá, lượt đánh giá 5 sao là cao nhất, 1 sao là thấp nhất và số sao trung bình cho mỗi đánh giá là xấp xỉ 3,5 sao.





Hình 4.29. Đếm số lượng từng sao trong cột ‘rating\_star’ và sự phân bổ

Có thể thấy rằng số lượt đánh giá 5 sao chiếm số lượng lớn là 5333 lượt, đồng thời số lượt đánh giá từ 3 đến 5 sao chiếm phần lớn trong bộ dữ liệu với 75,5% và số lượt đánh giá 2 sao chiếm số lượng ít nhất với 10,7%.

* **Xoá biểu tượng cảm xúc có định dạng <emoji>...</emoji>**

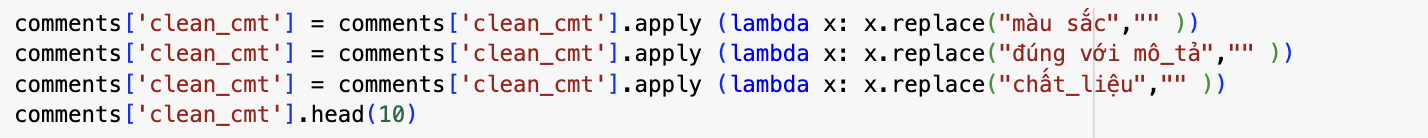
Khi tiền xử lý dữ liệu, nhóm đã chuyển emoji thành định dạng ‘<emoji> </emoji>’, tuy nhiên khi trực quan hoá dữ liệu, nhóm sẽ tiến hành bỏ dạng chữ của emoji đi để không làm ảnh hưởng đến những dữ liệu dạng chữ khác trong phần đánh giá của khách hàng.



Hình 4.30. Hàm xử lý xoá emoji có định dạng <emoji> </emoji>

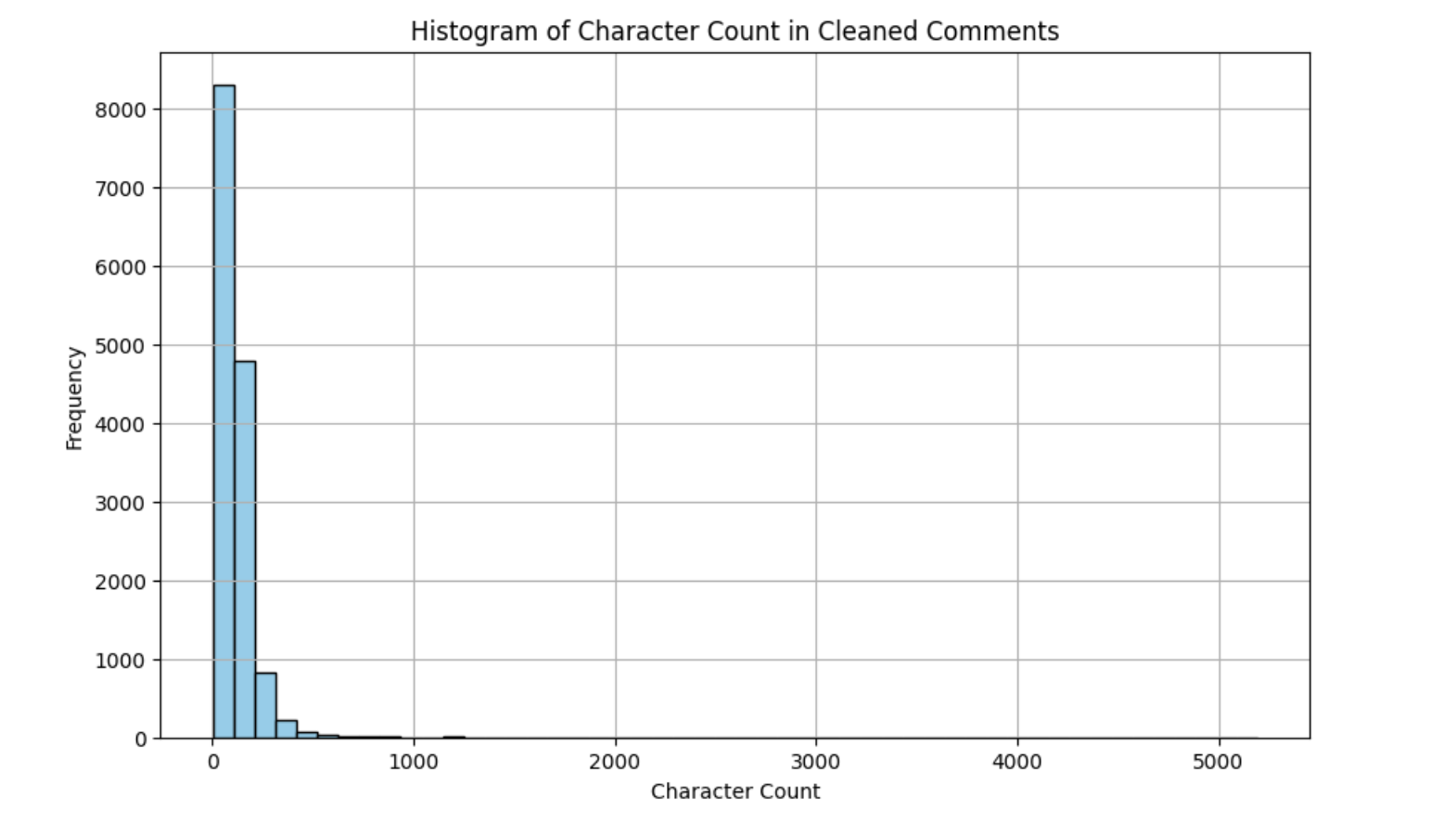
* **Xoá dữ liệu mặc định có sẵn trong phần đánh giá**

Các bình luận của shopee thường sẽ mặc định các mục như là "đúng với mô tả", "màu sắc", "chất liệu" để khách hàng có thể dễ dàng đánh giá theo các mục này, tuy nhiên, các dữ liệu này có thể làm nhiễu quá trình trực quan hoá dữ liệu và nhóm chỉ muốn phân tích những đánh giá khách hàng tự nhập vào, vì vậy sẽ tiến hành xoá đi những dữ liệu mặc định có sẵn này



Hình 4.31. Xoá những dữ liệu mặc định có sẵn trong phần đánh giá

* **Trực quan hoá cột dữ liệu ‘clean\_cmt’**
* **Đếm số lượng ký tự có trong từng đánh giá**

****

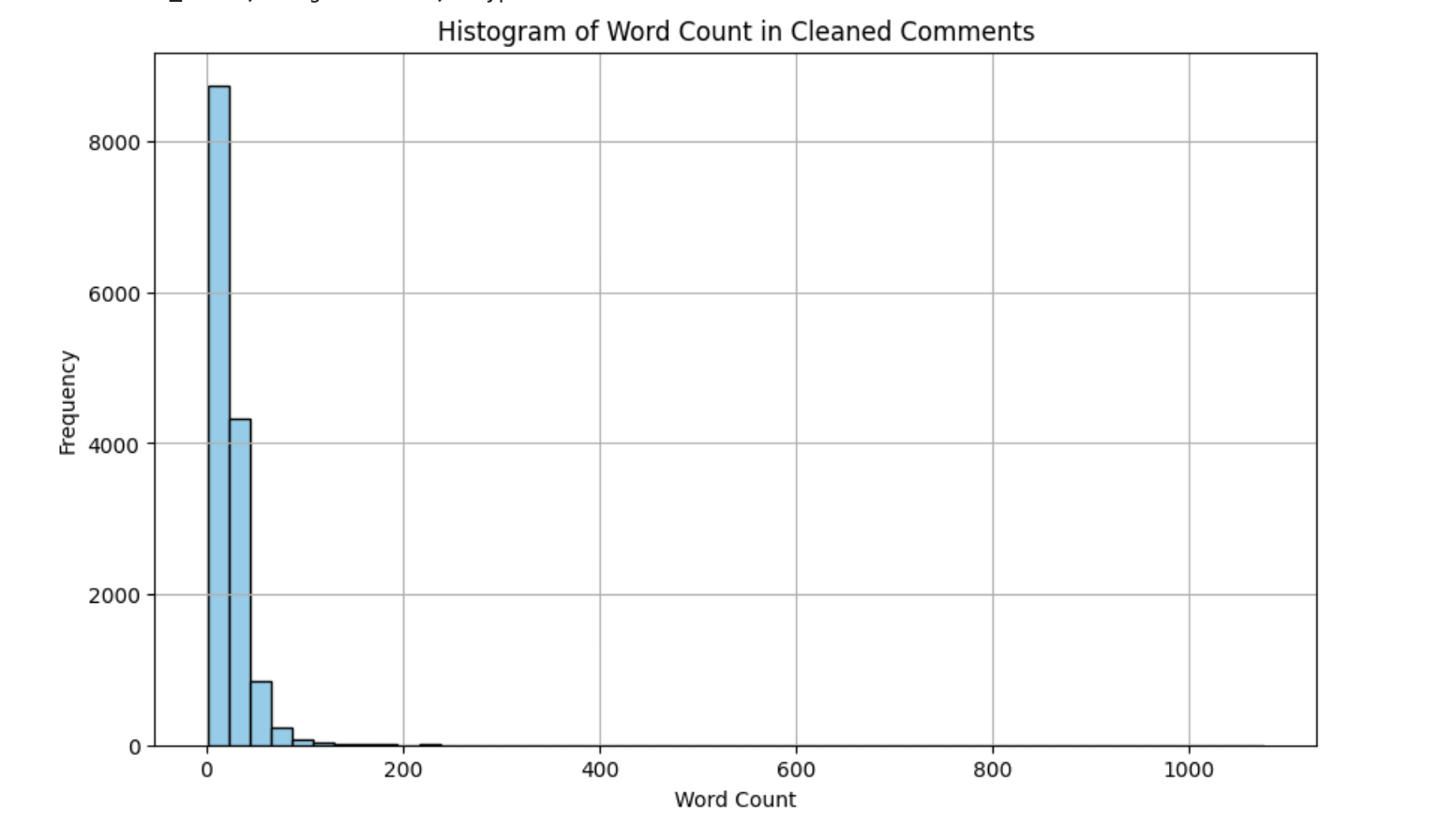
Hình 4.32. Biểu đồ histogram thể hiện số lượng ký tự có trong từng đánh giá



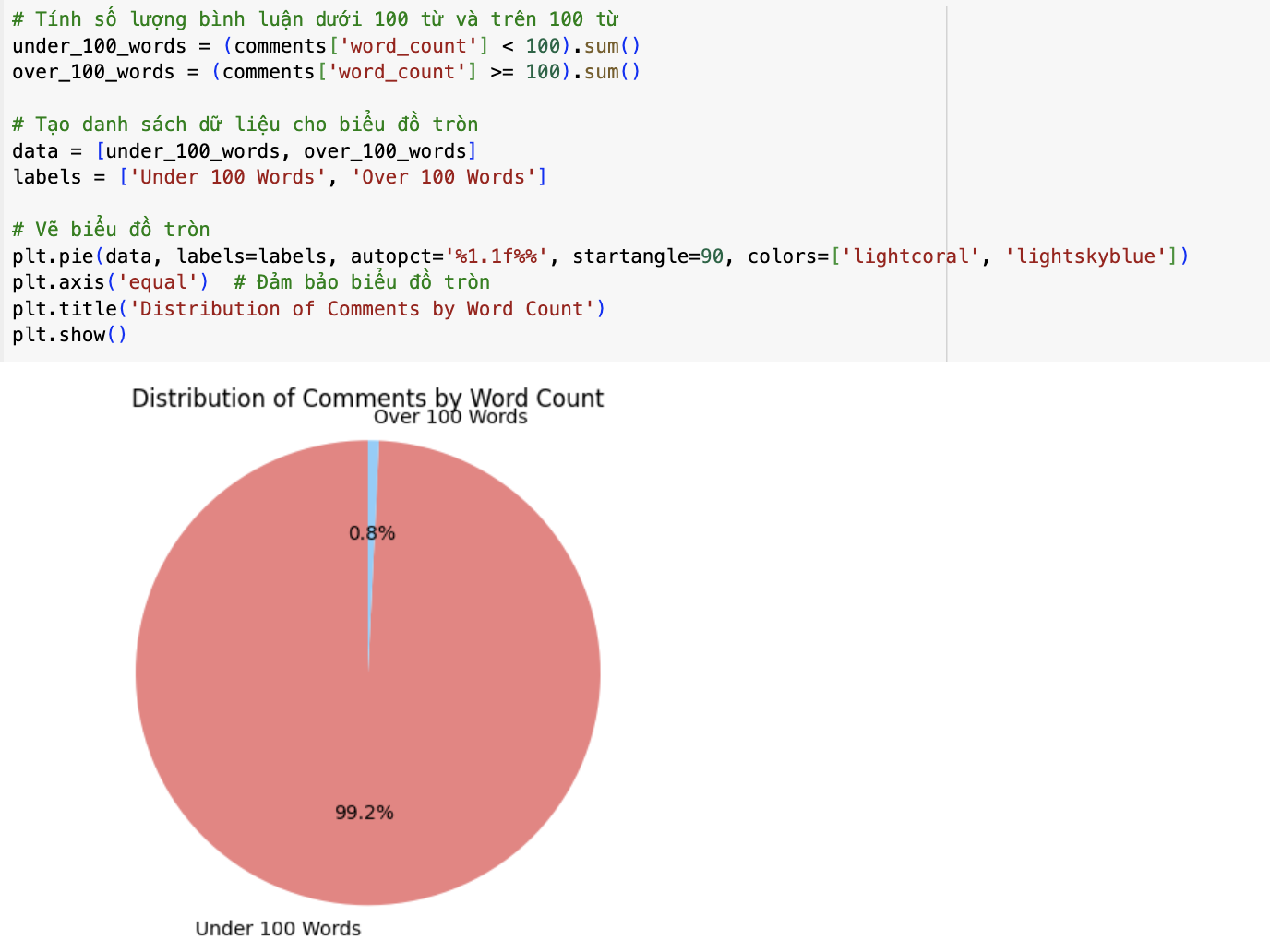
Hình 4.33. Số lượng ký tự trung bình

Độ dài của các bình luận phân bổ rơi vào khoảng 0-400 ký tự chiếm phần lớn, với độ dài trung bình là khoảng 115 ký tự và rất ít bình luận có trên 500 ký tự.

* **Đếm số lượng từ có trong từng đánh giá**



Hình 4.34. Biểu đồ histogram thể hiện số từ có trong từng đánh giá



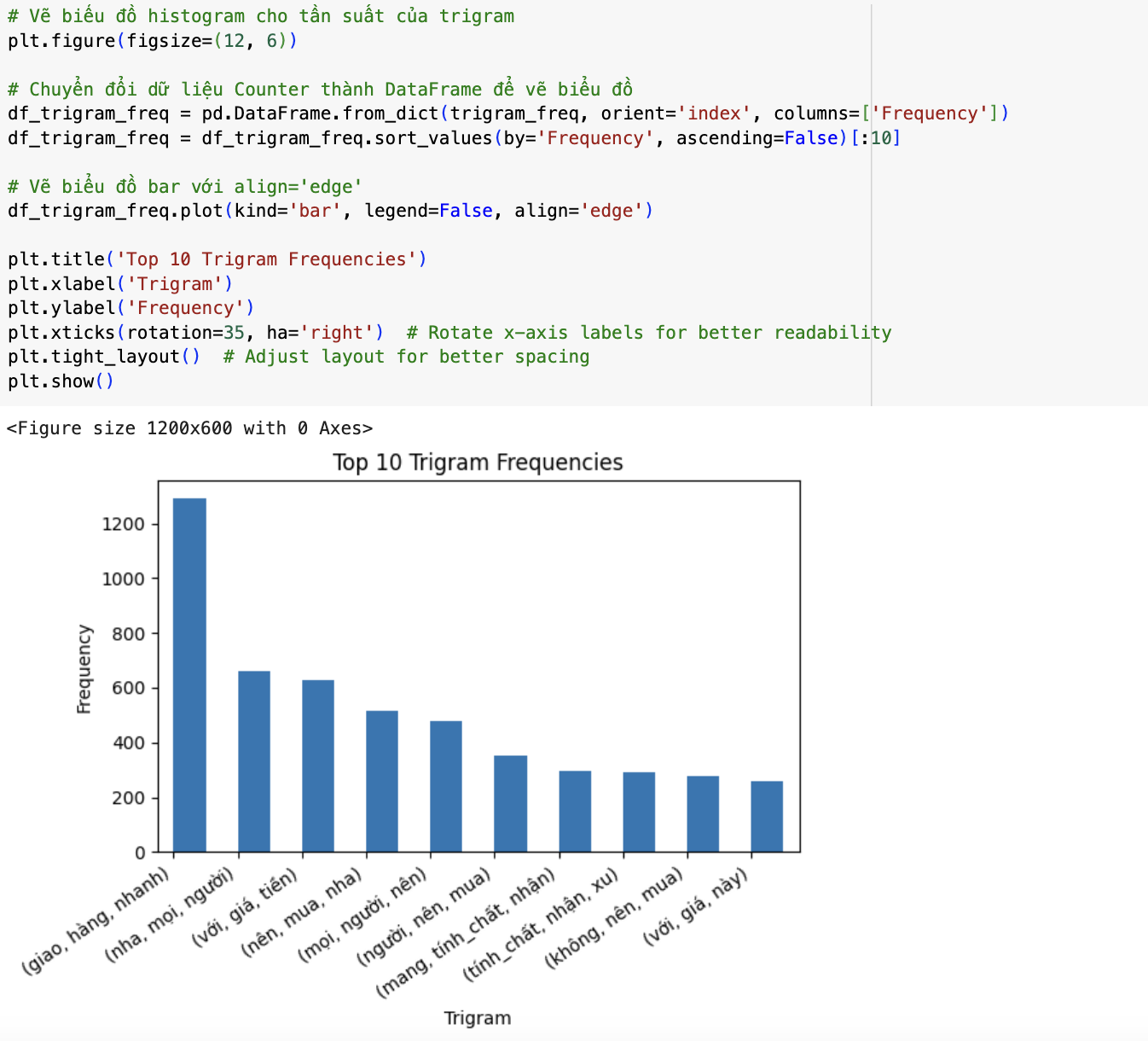
Hình 4.35. Sự phân bổ số lượng đánh giá có trên 100 từ và dưới 100 từ

Dựa vào Hình 4.34 và Hình 4.35 có thể thấy rằng đa số các đánh giá nằm trong khoảng từ 0 - 100 từ chiếm đến 99.2%, các bình luận trên 100 từ chỉ chiếm 0,8%.

* **Tokenize và tạo trigram cho mỗi đánh giá**



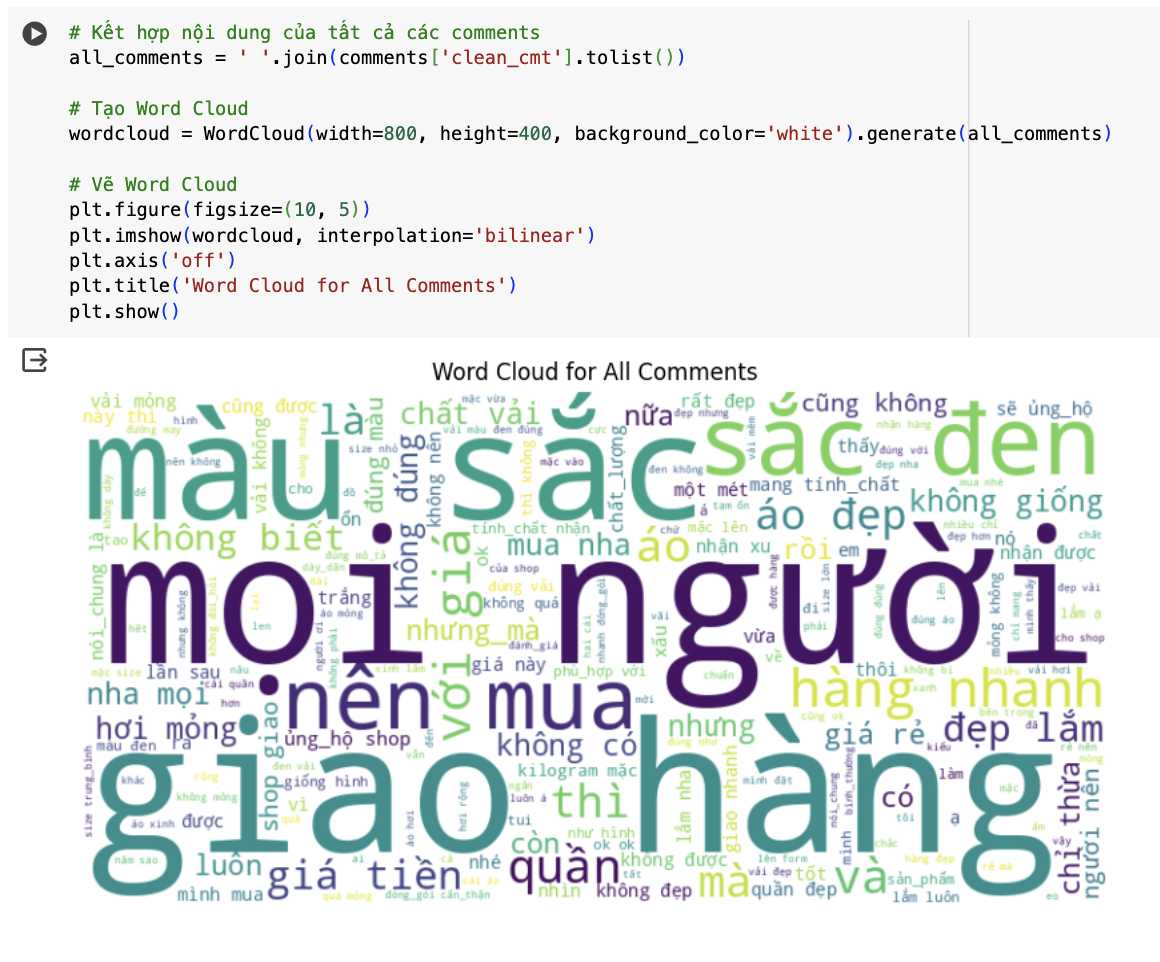
Hình 4.36. Tokenize và tạo trigram cho mỗi đánh giá



Hình 4.37. Biểu đồ thể hiện tần suất xuất hiện của 10 triagram phổ biến nhất

Các cụm từ xuất hiện nhiều nhất là cụm "giao hàng nhanh" với tần suất rơi vào khoảng 1300 lần, các cụm từ còn lại như "nên mua nha", "nha mọi người",... rơi vào tần suất xuất hiện từ 300-700 lần

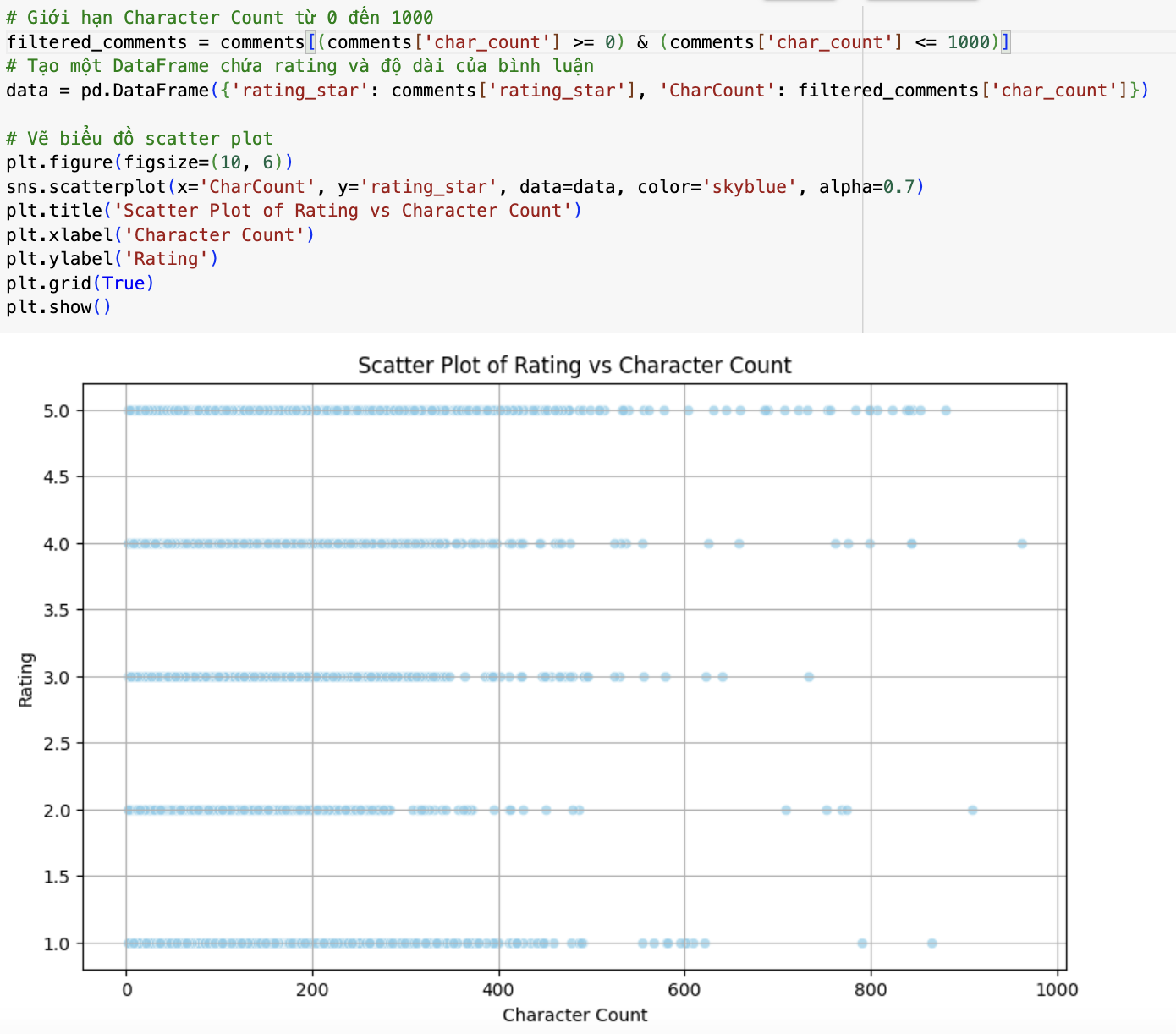
* **WordCloud**

****

Hình 4.38. WordCloud

Quan sát biểu đồ trên, giá trị 'giao hàng', "mọi người" có tần suất lặp lại nhiều nhất nên có kích thước lớn nhất với xanh lá trên biểu đồ. Tiếp theo đó là các giá trị như: 'nên mua', 'hàng nhanh',... Các giá trị có tần suất lặp lại càng ít thì màu sắc càng mờ và cỡ chữ càng nhỏ như: 'quá mỏng', 'chỉ thừa', ‘màu sắc',...

* **Sự tương quan giữa ‘rating\_star' và ‘clean\_cmt'**



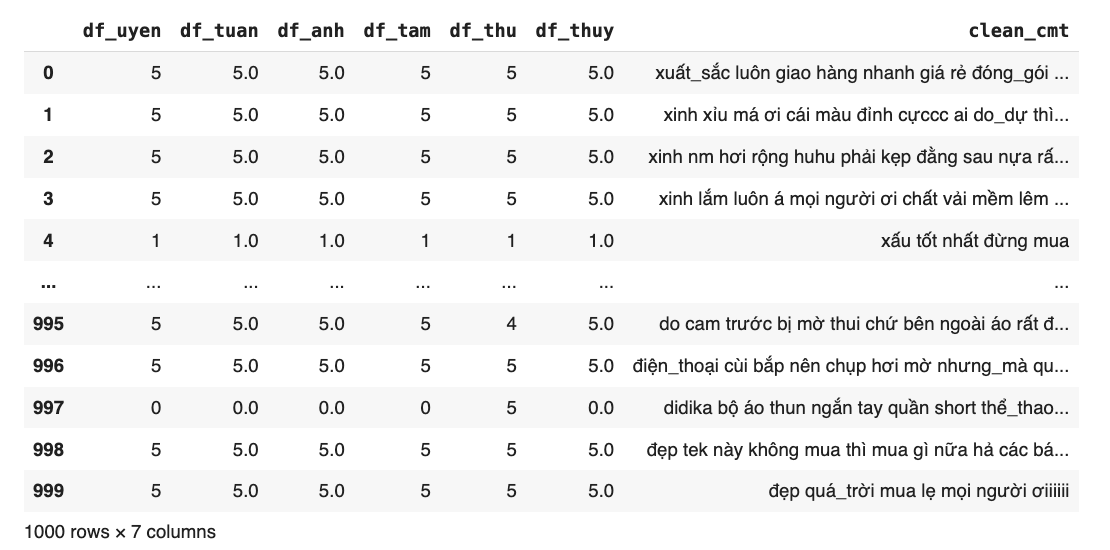
Hình 4.39. Biểu thị sự tương quan giữa ‘rating\_star' và ‘clean\_cmt'

Nhóm thực hiện vẽ biểu đồ thể hiện sự tương quan giữa rating với độ dài ngắn của bình luận, tuy nhiên vì các comment sẽ rơi vào độ dài từ 0-1000 là chủ yếu nên nhóm sẽ giới hạn để có thể trực quan dữ liệu rõ hơn. Theo đó có thể thấy các bình luận càng dài sẽ có rating càng cao rơi vào 4-5 sao, các bình luận 1-3 sao có độ dài ít hơn

## Chuẩn bị mô hình

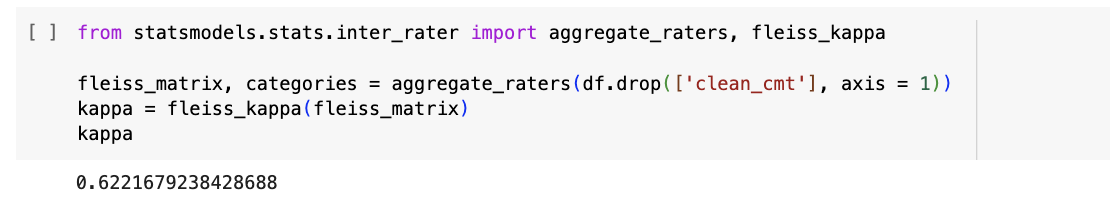
* **Chuẩn bị bộ dữ liệu huấn luyện**

Gắn nhãn là một phần quan trọng trong quá trình huấn luyện mô hình trong đề tài này của nhóm, mặc dù có các kỹ thuật gắn nhãn tự động được thiết lập tốt nhưng những kỹ thuật tự động này vẫn không cung cấp nhãn chính xác và đáng tin cậy cho mọi trường hợp sử dụng máy học trong nhiều ngành [Teodor Fredriksson, 2020], đặc biệt là ngôn ngữ phân tích trong đề tài này là Tiếng Việt với ngành hàng thời trang thì chưa có công cụ gắn nhãn chính xác và đúng mục đích của đề tài. Như đã trình bày ở trên, nhóm sẽ đề xuất phân loại và từ đó đánh giá điểm bình luận do vậy trước khi thực hiện huấn luyện mô hình nhóm đã tiến hành gắn nhãn 1000 dòng dữ liệu đầu tiên với 5 nhãn là số sao tương tứng “1”, “2”, “3”, “4”, “5” theo những quy tắc mà nhóm đưa ra để đánh giá độ được độ điểm đánh giá một cách chính xác nhất. Trong đề tài lần này là **“Đề xuất phân loại và điểm đánh giá cho bình luận của ngành hàng thời trang trên sàn thương mại điện tử Shopee”**, nhóm đánh giá các bình luận có liên quan đến vấn đề giao hàng hay người giao hàng sẽ không phản ảnh được độ tin cậy của sản phẩm, do vậy với những bình luận có những vấn đề liên quan đến giao hàng nhóm sẽ xem xét và gắn nhãn với số sao là “0” thể hiện bình luận này không dùng để đánh giá sản phẩm mà để đánh giá những vấn đề liên quan khác nhằm không ảnh hưởng đến số sao thực sự của sản phẩm và khi tiến hành train mô hình để dự đoán số sao chúng tôi sẽ lược bỏ những bình luận có số sao bằng “0” đã gắn thẻ này nhằm không ảnh hưởng đến độ chính xác của mô hình.



Hình 4.40. Gắn nhãn 1000 dòng dữ liệu đầu tiên

Khi thực hiện gán nhán 1000 dòng đầu tiên, nhóm sử dụng Fleiss Kappa để đo độ đồng thuận của 6 thành viên tham gia gắn nhãn, kết quả thu được độ đồng thuận là 62,21%.



Hình 4.41. Sử dụng Fleiss Kappa để đo độ chính xác gắn nhãn dữ liệu

Khi thực hiện gắn nhãn bộ dữ liệu với 6 nhãn, nhóm nghiên cứu đã đạt được độ đồng thuận là 62.21%, đây là một độ đồng thuận ở mức tốt, do vậy nhóm tiếp tục triển khai gắn nhãn những dòng dữ liệu còn lại trong tập dữ liệu để chuẩn bị cho việc huấn luyện mô hình.

Khác với tiếng Anh, đặc trưng của ngôn ngữ tiếng Việt là từ ghép, do đó nhóm tiến hành huấn luyện mô hình trên 2 tập dữ liệu là tập dữ liệu có tách từ (word\_segment) và không có tách từ để đo độ hiệu quả của mô hình.

* Cột ‘content’ là cột chứa các bình luận không được tách từ
* Cột ‘clean\_cmt’ là cột chứa các bình luận được tách từ

Tiếp theo, nhóm đã thực hiện một phương pháp gọi là "data augmentation" (tăng cường dữ liệu) để gia tăng số lượng dữ liệu huấn luyện. Cụ thể, nhóm đã sử dụng phương pháp này để tăng cường dữ liệu bằng cách chuyển đổi các dòng dữ liệu từ ngôn ngữ gốc (Tiếng Việt) sang một ngôn ngữ khác (Tiếng Anh), sau đó chuyển lại thành ngôn ngữ gốc. Kết quả là, số lượng dòng dữ liệu đã tăng lên gấp đôi, từ 14.000 dòng ban đầu lên 28.000 dòng, tạo điều kiện thuận lợi cho quá trình huấn luyện mô hình.

### Machine Learning

Trong phần Machine Learning nhóm sử dụng 3 mô hình máy học là Naive Bayes, Logistic Regression, và Linear Support Vector Machine (SVM) để thực hiện huấn luyện mô hình. Có rất nhiều thuật toán học máy được sử dụng để giải quyết cho bài toán phân lớp văn bản (Text Classification) tương tự với đề tài của nhóm nghiên cứu. Tuy nhiên, Naive Bayes là thuật toán có thời gian chạy nhanh và độ chính xác cao nên thường được sử dụng cho các bài toán phân lớp văn bản. Trong một nghiên cứu của tác giả Đặng Văn Nam (2020) trình bày cụ thể việc xây dựng một mô hình học máy với thuật toán Naive Bayes sử dụng đặc trưng TF-IDF (Term Frequency – Inverse DocumentFrequency) trong phân lớp văn bản đã cho thấy hiệu quả của thuật toán Naive Bayes so với các thuật toán học máy khác với thời gian huấn luyện, kiểm thử mô hình rất nhanh và độ chính xác cao. Hồi quy Logistic (Logistic Regression) là thuật toán Supervised Learning (học máy có giám sát), là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của một yếu tố dựa trên yếu tố còn lại. Trong nghiên cứu của nhóm tác giả Priyanshi Kathuria, Parth Sethi, Rithwick Negi (2022) về bình luận và rating trên các sàn thương mại điện tử, kết quả dự đoán cho thấy với mô hình Logistic Regression cho kết quả có độ chính xác cao nhất ở cả hai nhóm phân tích là bình luận (88,18%) và đánh giá (80,68%). Mô hình SVM là một mô hình có thể sử dụng cho bài toán phân lớp (classification) cũng như bài toán hồi quy (regression). Đây là thuật toán hỗ trợ phân loại rất phổ biến và hiệu quả, có thể áp dụng trong học có giám sát hoặc bán giám sát. Trong nghiên cứu của tác giả Đặng Quang Vinh (2023) đã xem xét bài toán định giá quyền chọn sử dụng công thức Black-Scholes bằng một số thuật toán máy học có giám sát (supervised machine learning). Kết quả thực nghiệm cho thấy mô hình SVM có nhiều khả năng ước lượng giá quyền chọn với độ chính xác cao so với các mô hình truyền thống khác.

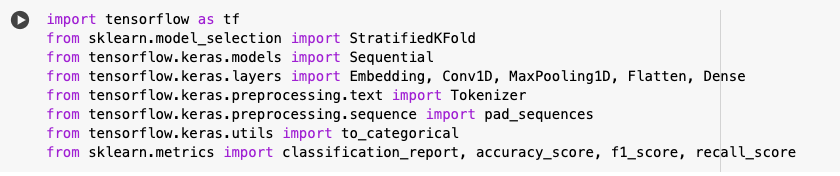
Vì vậy, nhóm sẽ thực hiện Machine Learning với 3 mô hình Naive Bayes, Logistic Regression, và Linear Support Vector Machine (SVM)

****

Hình 4.42. Chuẩn bị hàm Machine Learning

### Deep Learning

Đối với Deep learning, nhóm đã sử dụng mô hình CNN để huấn luyện và đánh giá mô hình dựa trên tập dữ liệu được chia thành nhiều fold để đánh giá khả năng tổng quát hóa của mô hình. Trong một nghiên cứu của tác giả Thái Thuận Thương (2021) đã chứng minh được mô hình CNN có độ chính xác cao hơn mô hình Vanilla và DNN. Do vậy nhóm đã viết hàm để sử dụng mô hình CNN như sau:





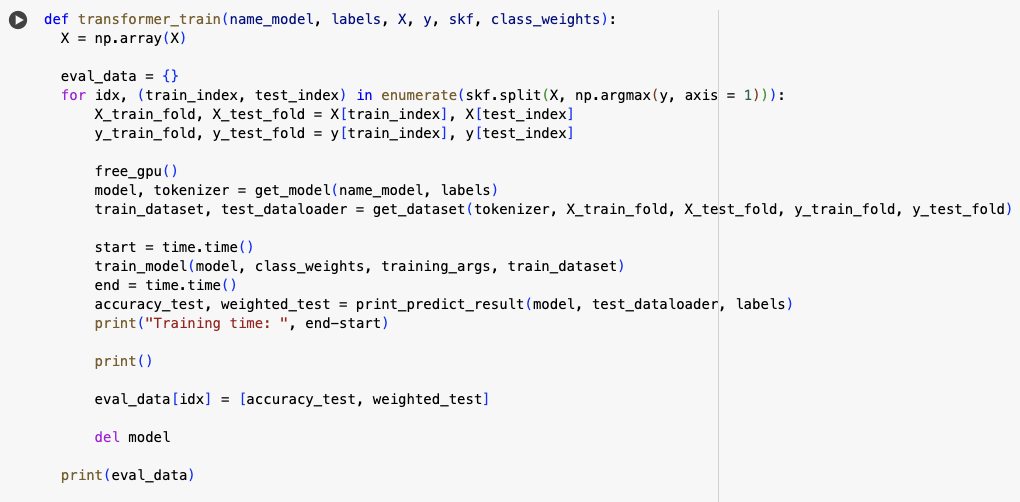
Hình 4.43. Chuẩn bị hàm Deep Learning

### Transformer

Mô hình được nhóm sử dụng ở Transformer là "vinai/phobert-base-v2", là một biến thể của mô hình phoBERT. PhoBERT là mô hình được huấn luyện riêng dành cho tiếng Việt, đã có nghiên cứu gần đây cho thấy rằng mô hình phoBERT hoạt động tốt hơn nhiều và đem lại hiệu quả cao so với BERT và FastText (Ngô Văn Sơn và cộng sự, 2022).

Trong mô hình này nhóm nghiên cứu chọn số lần huấn luyện mô hình (epochs) là 5. Có thể đây là một con số khá nhỏ, tuy nhiên nhóm đã thực hiện thử với số lần huấn luyện mô hình là 30 thì kết quả đưa ra đã chỉ ra rằng hệ số loss sau đó không giảm và hầu như là ngang bằng, không ảnh hưởng đáng kể đến độ chính xác của kết quả huấn luyện mô hình. Vì vậy, nhóm đã lựa chọn epochs bằng 5 để tiết kiệm tài nguyên trong quá trình huấn luyện mà vẫn cho ra được kết quả tối ưu.

​Do vậy nhóm đã viết hàm để sử dụng mô hình phoBert như sau: ​



Hình 4.44. Chuẩn bị hàm Transformer

|  |  |
| --- | --- |
| **Các tham số trong mô hình** | **phoBert** |
| **Epochs** | 5 |
| **Batch size** | 64 |
| **Learning Rate** | 1e - 4 |
| **K** | 5 |
| **Max Length** | 256 |
| **Weight\_decay** | 0.01 |

Bảng 4.1. Bảng thông số trong mô hình phoBert

## Kết quả phân tích

### In-Out Domain

#### No Token

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** | **Recall** |
| MultinomialNB | 0.782502 | 0.793012 | 0.782512 |
| Logistic Regression | 0.914971 | 0.913178 | 0.914969 |
| LinearSVC | 0.927936 | 0.929061 | 0.929048 |
| Deep Learning | 0.9257172 | 0.9263021 | 0.92571722 |

Bảng 4.2. Kết quả huấn luyện in-out domain trên tập dữ liệu no-token

Đối với tập dữ liệu no token thì qua quá trình huấn luyện in-out domain, kết quả thu được cho thấy cả ba mô hình Machine Learning bao gồm: Multinomial Naive Bayes, Logistic Regression và Linear Support Vector Classification đều cho kết quả tốt và không có sự chênh lệch quá nhiều giữa 3 chỉ số Accuracy, F1 Score và Recall của mỗi mô hình. Tuy nhiên, trong khi Logistic Regression và Linear SVC cho độ chính xác cùng bộ chỉ số F1 Score và Recall cao, với accuracy lên đến 0.9149 - 0.9279, thì mô hình Naive Bayes có độ chính xác thấp hơn so với hai mô hình trên, nhưng vẫn đạt mức tương đối (0.7825). Điều này cho thấy rằng các mô hình học máy với tiêu biểu là 3 đại diện trên có khả năng tốt trong việc phân loại bình luận, đánh giá của khách hàng, đặc biệt là hai mô hình Linear và Logistic Regression.

Mô hình Deep Learning CNN cũng gây ấn tượng với độ chính xác 0.9257 cùng F1 Score và Recall cao tương ứng. Chứng minh rằng đây là một mô hình tốt cho việc phân loại dữ liệu đánh giá của khách hàng.

Mô hình học sâu (Deep Learning) đã đạt được hiệu suất rất tốt trong các trường hợp, tuy nhiên, hiệu suất của nó không phải là cao nhất nhưng lại cần nhiều thời gian và tài nguyên hơn để huấn luyện hơn so với 3 mô hình học máy. LinearSVC là mô hình mang lại độ chính xác cao nhất trong bộ 4 mô hình trên khi huấn luyện dữ liệu in-out domain với cả 2 bộ dữ liệu token và token. Và ở chiều ngược lại, mô hình có bộ chỉ số kém khả quan nhất là Multinomial Naive Bayes.

Đối với mô hình phoBERT, các kết quả này cho thấy mô hình đã đạt được hiệu suất rất tốt trên cả hai lớp "in" và "out", với độ chính xác và các chỉ số F1 Score rất cao lên đến 0.963. Bên cạnh đó ta cũng có thể thấy độ chính xác trên lớp dữ liệu in domain (liên quan đến sản phẩm) có độ chính xác cao hơn so với lớp out. Điều này có thể được lý giải bởi số lượng dữ liệu đầu vào của lớp “in” cao hơn 4 lần so với lớp “out”. Tuy vậy, một phần lớn dữ liệu trong lớp out domain có chứa những từ khóa đặc thù như “shipper”, “giao hàng”, “thái độ của shop”... nên dù dữ liệu ít hơn vẫn cho ra một kết quả tốt và không kém hơn quá nhiều so với lớp in domain.

Qua những phân tích trên, ta có thể kết luận rằng mô hình transformer phoBERT với đặc thù bộ dữ liệu được xây dựng dành riêng cho tiếng Việt giúp thu được kết quả độ chính xác cao nhất trong cả 5 mô hình.

#### Token

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** | **Recall** |
| MultinomialNB | 0.775645 | 0.784400 | 0.775658 |
| Logistic Regression | 0.917934 | 0.916757 | 0.917933 |
| LinearSVC | 0.930532 | 0.931689 | 0.932012 |
| Deep Learning | 0.927569 | 0.927325 | 0.927569 |

Bảng 4.3. Kết quả huấn luyện in-out domain trên tập dữ liệu token

Dựa trên bảng kết quả được cung cấp, có thể thấy cả ba mô hình Machine Learning bao gồm Multinomial Naive Bayes, Logistic Regression và Linear Support Vector Classification đều đạt được độ chính xác cao trên tập dữ liệu token.

Mô hình Deep Learning CNN cũng gây ấn tượng với độ chính xác 0.9275 cùng F1 Score và Recall cao tương ứng. Chứng minh rằng đây là một mô hình tốt cho việc phân loại dữ liệu đánh giá của khách hàng.

Tuy nhiên, ta có thể thấy Linear Support Vector Classification có độ chính xác cao nhất (0.930532) Điều này cho thấy Linear Support Vector Classification có khả năng phân loại chính xác các token tốt hơn so với hai mô hình còn lại. Tương tự như độ chính xác, Linear Support Vector Classification cũng có F1 Score cao nhất (0.931689). F1 Score là một chỉ số đánh giá hiệu suất tổng hợp của mô hình, tính đến cả độ chính xác và độ nhạy bén. Do đó, kết quả này cho thấy Linear Support Vector Classification có khả năng cân bằng tốt giữa hai yếu tố này hơn so với hai mô hình còn lại.Linear Support Vector Classification cũng có Recall cao nhất (0.932012). Recall là một chỉ số đánh giá khả năng xác định chính xác các token thuộc lớp tích cực. Do đó, kết quả này cho thấy Linear Support Vector Classification có khả năng nhận diện các token thuộc lớp tích cực tốt hơn so với ba mô hình còn lại.

Mô hình phoBERT đã thể hiện hiệu suất ấn tượng trong phân loại đánh giá sản phẩm tiếng Việt, đạt độ chính xác accuracy là 0.963 và F1 Score lên đến 0.98 cho lớp "in". Lớp "in" (liên quan sản phẩm) có độ chính xác cao hơn do lượng dữ liệu đầu vào lớn, trong khi lớp "out" (không liên quan sản phẩm) vẫn đạt kết quả tốt nhờ sở hữu các từ khóa đặc thù. So sánh với 4 mô hình khác, phoBERT khẳng định vị trí dẫn đầu nhờ bộ dữ liệu tiếng Việt chuyên biệt.

Kết quả này cho thấy phoBERT là lựa chọn tối ưu cho việc phân loại đánh giá sản phẩm tiếng Việt, mang đến khả năng phân biệt chính xác và hiệu quả cao. Việc tiếp tục thu thập dữ liệu và ứng dụng phoBERT vào các bài toán khác liên quan đến đánh giá sản phẩm và dịch vụ hứa hẹn mở ra nhiều tiềm năng phát triển trong tương lai.

**Đánh giá**

Căn cứ trên kết quả sau khi huấn luyện mô hình, có thể thấy rằng đối với việc train dữ liệu để dự đoán phân loại bình luận đánh giá có thật sự liên quan đến sản phẩm (in-out domain) thì bộ dữ liệu token cho hiệu suất tốt hơn ở đa số các mô hình học máy và học sâu (deep learning) so với bộ no-token. Tokenization đã giúp các mô hình hiểu và xử lý văn bản một cách hiệu quả hơn, từ đó giúp cải thiện hiệu suất huấn luyện và phân loại dữ liệu .

### Predict Star

#### No Token

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** | **Recall** |
| MultinomialNB | 0.289081 | 0.223858 | 0.289081 |
| Logistic Regression | 0.619061 | 0.608296 | 0.619061 |
| LinearSVC | 0.616894 | 0.604787 | 0.619873 |
| Deep Learning | 0.616468 | 0.606878 | 0.616468 |

Bảng 4.4. Kết quả huấn luyện dự đoán sao (predict star) trên tập dữ liệu no-token

Dựa vào kết quả sau khi huấn luyện mô hình dự đoán số sao với tập dữ liệu no token, có thể thấy rằng mô hình Naive Bayes đem lại độ chính xác thấp nhất. Điều này có thể là do mô hình Naive Bayes không xử lý tốt các đặc trưng liên tục hoặc tương đối lớn như làm việc trực tiếp với dữ liệu văn bản.

Logistic Regression và Linear SVC thường mang lại hiệu suất tốt hơn, có thể do khả năng của chúng trong việc tìm ra mối quan hệ tuyến tính giữa các đặc trưng và nhãn. Điều này phản ánh rõ trong kết quả với độ chính xác và các chỉ số đánh giá khác như F1-score và Recall không có sự chênh lệch đáng kể giữa hai mô hình này.

Mô hình Deep Learning trong tập dữ liệu no token đã cho thấy khả năng dự đoán tốt, gần bằng với các mô hình truyền thống như Logistic Regression và Linear SVC. Điều này ngụ ý rằng mô hình đã học được các mẫu và mối quan hệ phức tạp trong dữ liệu mà không cần sự tiền xử lý từ việc token. Tuy nhiên, dữ liệu ở đây đa số là các từ ngắn và đơn giản, với các loại dữ liệu phức tạp hơn cần lưu ý đến yếu tố về độ phức tạp của mô hình và tài nguyên cần thiết cho việc huấn luyện.

Mô hình phoBERT đối mặt với vấn đề mất cân bằng trong dữ liệu đầu vào. Dự đoán cho nhóm 4 sao có precision 0.91. Điều này có thể được lý giải bởi sự phong phú hơn của dữ liệu cho nhóm này. Mặt khác, dù có ít dữ liệu hơn, dự đoán cho nhóm 5 sao vẫn đạt được độ chính xác tương đối cao. Điều này là do các dữ liệu của nhóm 5 sao thường có các đặc điểm khác biệt so với các nhóm sao còn lại. Tuy nhiên dự đoán cho nhóm 1, 2 và 3 sao chưa đạt được kết quả tốt có thể là do dữ liệu đầu vào còn ít và phức tạp.

Đánh giá tổng quan qua các chỉ số độ chính xác, Recall và F1-score, mô hình phoBERT vẫn cho kết quả tốt nhất với độ chính xác là 0.756, cho thấy khả năng phân loại tốt nhất trong các mô hình đã được đánh giá.

#### Token

|  |  |  |  |
| --- | --- | --- | --- |
| **Mô hình** | **Accuracy** | **F1 Score** | **Recall** |
| MultinomialNaiveBayes | 0.331064 | 0.248430 | 0.331063 |
| LogisticRegression | 0.621885 | 0.610471 | 0.621885 |
| LinearSVC | 0.620995 | 0.608853 | 0.623123 |
| Deep Learning | 0.624825 | 0.619327 | 0.624825 |

Bảng 4.5. Kết quả huấn luyện dự đoán sao (predict star) trên tập dữ liệu token

Qua các kết quả sau khi huấn luyện dự đoán số sao với tập dữ liệu token, mô hình Naive Bayes cho kết quả độ chính xác thấp, còn Logistic Regression, SVM có độ chính xác khả quan hơn, cho thấy tiềm năng ứng dụng trong phân tích và phân loại bình luận. Kết quả giữa machine learning Logistic Regression, SVM và deep learning CNN có sự tương đồng. Mô hình CNN của deep learning có độ chính xác đạt 0.62, không chênh lệch giữa độ chính xác, Recall và F1 Score.

Đối với mô hình phoBERT, vì dữ liệu bị mất cân bằng giữa các sao nên kết quả mô hình cũng chịu ảnh hưởng. Dự đoán 4 sao có độ chính xác cao nhất, precision lên đến 0.92. Chỉ số F1 Score và Recall cũng cao hơn nhiều so với các sao khác. Điều này có thể được lí giải vì dữ liệu đầu vào của dự đoán 4 sao nhiều nhất (1880 dữ liệu), gấp đôi và gấp ba số dữ liệu của các sao còn lại. Tuy nhiên, đầu vào dự đoán 5 sao chỉ có 665 dữ liệu, nhưng lại có độ chính xác đạt 0.83. Nguyên nhân có thể do tính chất của tập dữ liệu cho các bình luận 5 sao có đặc điểm khác biệt so với các bình luận có số sao thấp hơn. Có thể những bình luận 5 sao thường có nội dung tích cực, ít hoặc thậm chí là không chứa các từ ngữ tiêu cực hoặc phản đối, dễ dàng để mô hình nhận biết. Trong khi đó, các bình luận có số sao thấp hơn có thể chứa nhiều phản hồi tiêu cực hoặc không hài lòng, điều này làm cho việc dự đoán khó khăn hơn.

Dữ liệu đầu vào của 1 sao là thấp nhất, chỉ có 664 dữ liệu, do đó độ chính xác cũng bị ảnh hưởng và thấp nhất trong các sao với precision là 0.62. Tương tự, 2 sao và 3 sao cũng có kết quả ở mức trung bình khá.

Các kết quả của mô hình phoBERT cho thấy mô hình có khả năng phân loại các bình luận 4 sao và 5 sao khá chính xác. Mặt khác mô hình lại không hiệu quả với các bình luận 1 sao, 2 sao và 3 sao, dễ dẫn đến việc phân loại nhầm các bình luận 1 sao, 2 sao và 3 sao.

Độ chính xác là thước đo quen thuộc trong đánh giá hiệu quả mô hình, tuy đơn giản và dễ hiểu nhưng nó lại thiếu đi tính toàn diện. Nó không phản ánh được khả năng phân biệt các lớp nhãn riêng lẻ của mô hình, cũng như bỏ qua những sai lệch tiềm ẩn trong trường hợp dữ liệu không cân bằng.

Do đó, để có cái nhìn đầy đủ và chính xác hơn về hiệu suất mô hình, cần phải có các chỉ số đánh giá khác bên cạnh độ chính xác như Recall và F1-score.

Sau khi đã đánh giá tổng quan qua ba chỉ số độ chính xác, Recall, F1-score, chúng tôi nhận thấy mô hình phoBERT cho kết quả tốt nhất với độ chính xác là 0.7562.

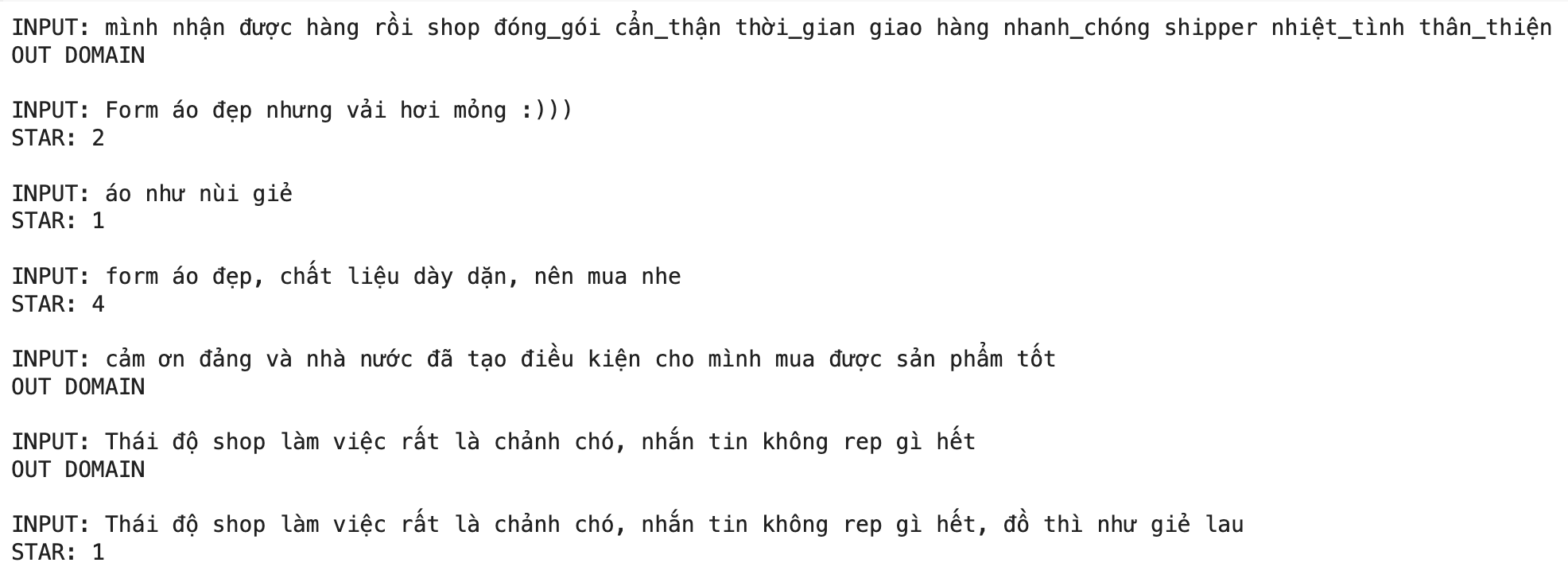
**Đánh giá**

So sánh kết quả giữa hai tập dữ liệu token và no token, cả độ chính xác, F1 Score và Recall đều không có sự khác biệt lớn. Kết quả đó không có nghĩa là việc token tập dữ liệu không có hiệu quả mà bởi vì đặc trưng của tập dữ liệu này là đánh giá sản phẩm, hầu hết các đánh giá có số lượng từ không nhiều nên việc token dữ liệu chưa thể hiện rõ được ưu thế của nó. Nếu các đánh giá có số lượng từ nhiều hơn, đánh giá của khách hàng càng chi tiết thì token dữ liệu sẽ càng chứng minh được ưu thế của mình và kết quả giữa token và no token sẽ có chênh lệch.

## Ứng dụng mô hình

Dựa vào các kết quả thực nghiệm, mô hình phoBERT trên tập dữ liệu token có hiệu suất tốt và đáng tin cậy trong việc phân loại bình luận và dự đoán số sao, do đó phoBERT trên tập dữ liệu token có thể được lựa chọn làm thuật toán phân loại và dự đoán ưu tiên cho bài toán này.

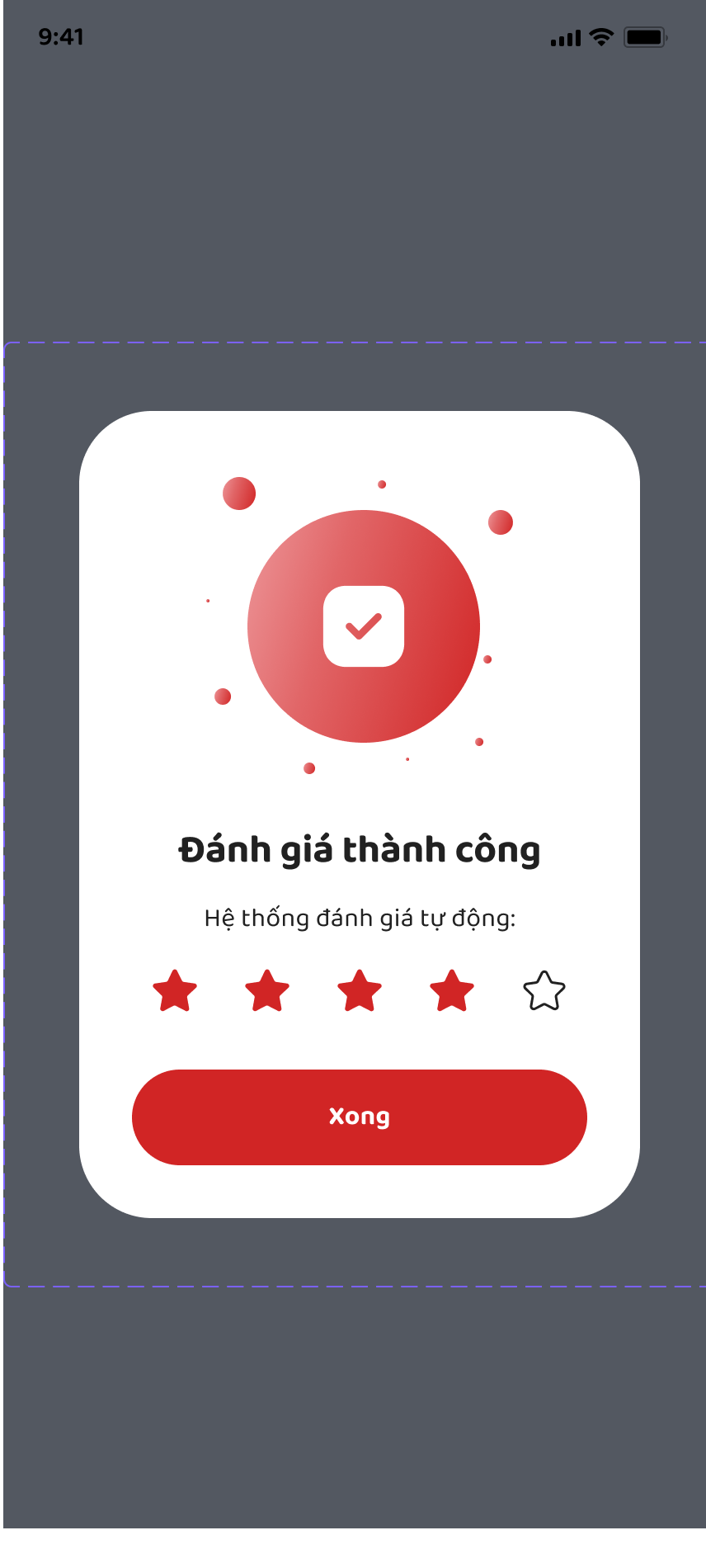
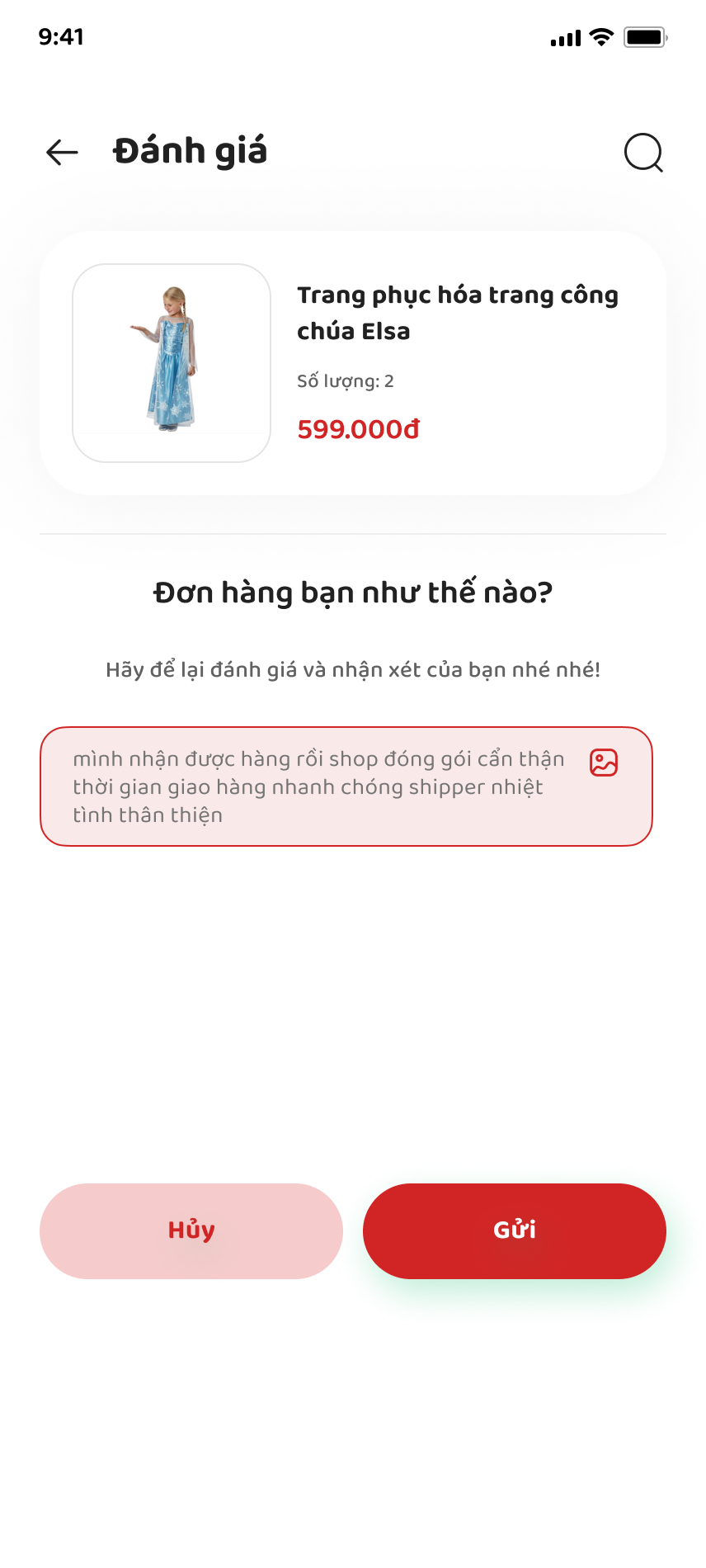
Nhóm sẽ ứng dụng mô hình này để có thể tự động phân loại các bình luận một cách hiệu quả, ngoài ra còn dự đoán điểm đánh giá (số sao) một cách chính xác, từ đó cung cấp cho người dùng một cái nhìn toàn diện để có thể đánh giá đúng hơn về chất lượng của sản phẩm. Từ đó, các người dùng khác cũng có cái nhìn hoàn thiện hơn về sản phẩm đó.



Hình 4.45. Kết quả ứng dụng mô hình

Hình 4.45 trên mô tả cách nhóm ứng dụng mô hình vào thực tế, khi khách hàng nhập bình luận của mình vào phần bình luận (INPUT) nếu bình luận đó đang được khách hàng sử dụng để đánh giá các vấn đề liên quan thì hệ thống sẽ tự động đánh giá vấn đề đó là “OUT DOMAIN” tức nghĩa là khách hàng đang không đánh giá chất lượng sản phẩm, tương tự như các bình luận còn lại, nếu khách hàng thực hiện đánh giá bình luận theo đúng quy tắc nhóm đã xây dựng, hệ thống sẽ đề xuất số sao tương ứng với bình luận đó.

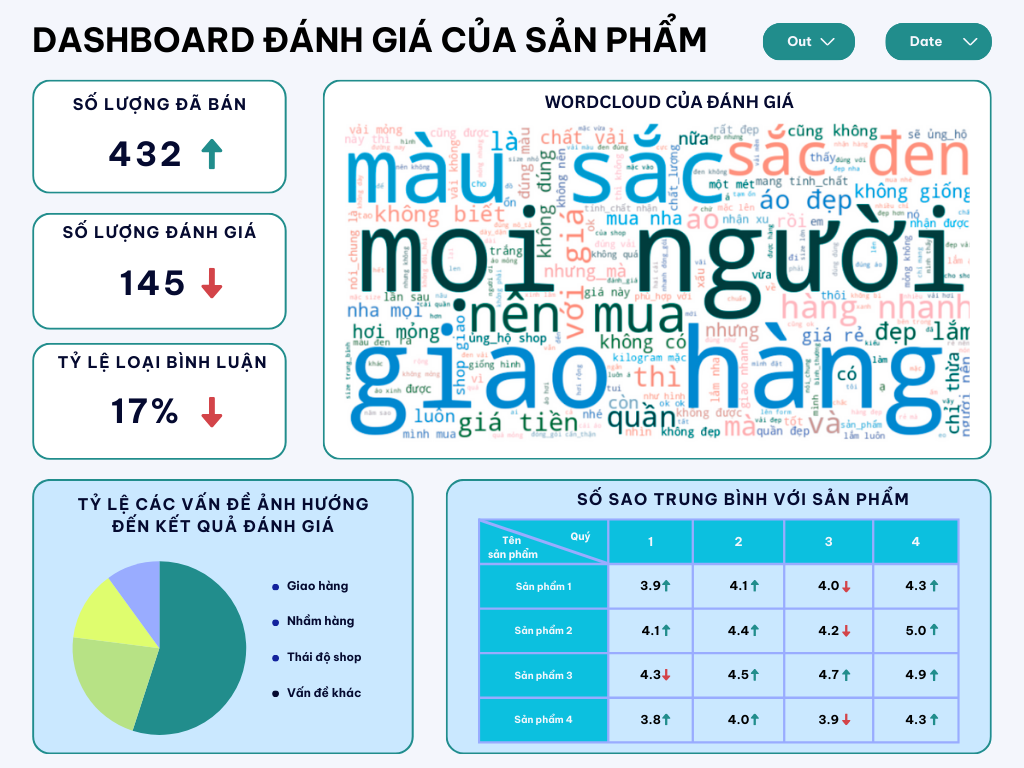
Một phần quan trọng của đề tài này là việc tạo ra các mockup để minh họa cho khách hàng cách hệ thống hoạt động. Điều này giúp người dùng dễ dàng hiểu và sử dụng hệ thống một cách linh hoạt và thuận tiện. Bằng cách này, nhóm hy vọng sẽ thúc đẩy sự chấp nhận và sử dụng của công nghệ trong lĩnh vực này, đồng thời nâng cao trải nghiệm mua sắm trực tuyến của người tiêu dùng.



Hình 4.46. Mockup ứng dụng để đề xuất phân loại và điểm đánh giá

Hình trên mô tả cách thức ứng dụng mô hình vào thực tế, khi khách hàng nhập đánh giá sản phẩm của mình và gửi đi, hệ thống sẽ tự động đánh giá liệu rằng sản phẩm này có đang đánh giá chất lượng sản phẩm không, hay đang đánh giá các dịch vụ liên quan. Khi khách hàng nhập đánh giá đúng với quy tắc đánh giá sản phẩm của doanh nghiệp ứng dụng mô hình thì hệ thống sẽ đề xuất số sao tương ứng với bình luận mà khách hàng nhập vào. Ngược lại, nếu bình luận của khách hàng đang không đánh giá chất lượng sản phẩm thì hệ thống sẽ thông báo đánh giá của khách hàng đang đánh giá các vấn đề liên quan, hệ thống sẽ ghi nhận lại đánh giá và gửi đến các bên liên quan để cải thiện dịch vụ chăm sóc khách hàng.

Bên cạnh đó, nhóm đề xuất thêm cách mà doanh nghiệp có thể ứng dụng kết quả nghiên cứu để xây dựng dashboard trực quan hoá số liệu nhằm dễ dàng phân tích được sản phẩm đang được đánh giá như thế nào.

Dashboard được xây dựng giúp doanh nghiệp có thể biết được các vấn đề đang ảnh hưởng đến kết quả đánh giá, tỷ lệ bình luận đánh giá đúng với quy tắc của doanh nghiệp hay các vấn đề liên quan khác như vấn đề giao hàng, thái độ làm việc của cửa hàng,...

Hình 4.47. Dashboard đánh giá của sản phẩm

# Kết luận và hướng phát triển

## Tóm tắt nội dung và kết quả nghiên cứu của đề tài

Với đề tài “Đề xuất phân loại và điểm đánh giá cho bình luận của ngành hàng thời trang trên sàn thương mại điện tử Shopee” nhóm đã vận dụng được kiến thức của môn học để lấy dữ liệu từ trang web, tiến hành các bước xử lý dữ liệu, trực quan hoá dữ liệu và chạy các mô hình học máy, học sâu và transformer bằng Python để hoàn thành tốt những mục tiêu đề ra ban đầu là ứng dụng các mô hình học máy, học sâu để đề xuất phân loại và điểm đánh giá cho bình luận cho ngành hàng thời trang của khách hàng khi mua sắm trên sàn thương mại điện tử Shopee.

Dựa trên kết quả đánh giá, chúng tôi nhận thấy rằng mô hình phoBERT trên tập dữ liệu token đã đạt được hiệu suất tốt nhất trong việc phân loại bình luận và cảm xúc khách hàng từ bình luận đánh giá sản phẩm thời trang nữ trên Shopee. Mô hình phoBERT trên tập dữ liệu token ở in-out domain đạt độ chính xác 0.963 và ở dự đoán số sao (predict star) đạt độ chính xác 0.7562. Các mô hình Linear Regression, SVM, CNN cũng cho kết quả khá với độ chính xác ở in-out domain là 0.91 - 0.93 và ở dự đoán số sao là khoảng 0.62.

Bên cạnh đó, nhóm đã ứng dụng được mô hình nghiên cứu cho tập dữ liệu khác ngoài tập dữ liệu huấn luyện ban đầu nhằm kiểm tra độ chính xác với những tập dữ liệu liên quan và thu được kết quả tốt.

Thông qua đề tài trên, nhóm thực hiện đã có thêm những kiến thức căn bản về phân tích dữ liệu bằng Python cũng như về ngành hàng mà nhóm thực hiện nghiên cứu, từ đó có thể phục vụ cho những đề tài sau này và công việc.

Tổng kết lại, các mô hình Linear Regression, SVM, CNN, phoBERT đều có ưu điểm và hạn chế riêng trong việc phân loại bình luận và dự đoán số sao. Để đạt được hiệu suất tốt hơn, cần tiếp tục nghiên cứu và cải thiện các mô hình phân tích cảm xúc để dự đoán số sao, đồng thời xem xét sự phù hợp với yêu cầu cụ thể của nghiên cứu.

## Ưu điểm

Trong quá trình lên ý tưởng và xây dựng bộ dữ liệu cho đề tài , nhóm đã đặt ra nhiều mục tiêu cụ thể theo từng giai đoạn khác nhau. Tất cả các mục tiêu ban đầu đều được thực hiện và hoàn thiện ở mức độ tốt. Đồng thời, nhóm cũng đạt được những thành quả ngoài dự kiến:

* Nhóm đã áp dụng linh hoạt các kiến thức được học vào đề tài, tìm hiểu và vận dụng kiến thức về ngôn ngữ Python để lấy dữ liệu từ trang web cũng như sử dụng đa dạng các mô hình Machine Learning, Deep Learning, Transfomer để tìm ra mô hình có độ chính xác cao nhất cho bộ dữ liệu mà nhóm thực hiện nghiên cứu.
* Sau khi dự đoán cho bộ dữ liệu thuộc ngành hàng thời trang mà nhóm nghiên cứu, nhóm có mở rộng dự đoán thêm cho bộ dữ liệu thuộc ngành hàng khác kiểm tra độ chính xác mô hình mà nhóm lựa chọn.
* Áp dụng các kiến thức của những môn học vào thực tế. Qua đó, nâng cao chất lượng đề tài.

## Nhược điểm

Bên cạnh những kết quả đã đạt được, đề tài vẫn tồn tại nhiều điểm hạn chế cần được khắc phục và hoàn thiện:

* Việc thực hiện gắn nhán dữ liệu để tiến hành chạy các mô hình khiến nhóm tốn nhiều thời gian để thực hiện.
* Bộ dữ liệu còn hạn chế và cùng nằm trong một ngành hàng là thời trang nên chưa có độ đa dạng về các đánh giá.
* Chưa phân tích dữ liệu trong thời gian thực để những nhà kinh doanh, nhà quản trị có thể ứng dụng để đưa ra các quyết định nhanh chóng.

## Hướng phát triển của đề tài

* Tiến hành thu thập thêm dữ liệu trong khoảng thời gian rộng hơn với nhiều web khác nhau để bộ dữ liệu có độ đa dạng hơn, đồng thời có thể thu thập những dữ liệu từ các ngành hàng có liên quan để có thể đưa ra những dự đoán về độ tin cậy của các đánh giá khách hàng một cách chính xác hơn.
* Trong tương lai, chúng tôi sẽ cải thiện quá trình thu thập dữ liệu liên tục theo thời gian thực, tối ưu hóa việc sửa lỗi chính tả, và cải thiện độ chính xác của việc gán nhãn tự động nhằm giảm thiểu sự can thiệp thủ công đối với dữ liệu ngày càng lớn hơn. Nghiên cứu sẽ tập trung vào việc cải thiện và điều chỉnh mô hình BERT để đạt được kết quả tốt hơn trên các tập dữ liệu phức tạp và lớn hơn.
* Bên cạnh việc thu thập, xử lý dữ liệu và đề xuất phân loại, đánh giá bình luận của khách hàng, nghiên cứu này còn tiến xa hơn khi đề xuất phương pháp tạo báo cáo trực quan về phân tích cảm xúc người dùng theo thời gian thực. Nhờ vậy, các nhà chiến lược kinh doanh có thể nắm bắt thông tin một cách nhanh chóng và kịp thời, từ đó đưa ra quyết định phù hợp để nâng cao mức độ hài lòng của khách hàng.

# Tài liệu tham khảo

[1] Loukili, M., Messaoudi, F., & El Ghazi, M. (2023). Sentiment Analysis of Product Reviews for E-Commerce Recommendation based on Machine Learning. International Journal of Advances in Soft Computing & Its Applications, 15(1).

[2] Sarulkar, O., Pitale, R., Tikhe, S., Giri, S., & More, R. (2023). Sentiment Analysis of Amazon Products Using a Hybrid Random Forest and Support Vector Machine Model. Available at SSRN 4618423.

[3] Gupta, K., Jiwani, N., & Afreen, N. (2023). A Combined Approach of Sentimental Analysis Using Machine Learning Techniques. Revue d'Intelligence Artificielle, 37(1).

[4] Mahmud, F. A. M., Mullick, S. B. R. A., & Anas, T. C. M. Sentiment Analysis of Women's Clothing Reviews on E-commerce Platforms: A Machine Learning Approach.

[5] Đạt, Đ. T. (2023). Ứng dụng phương pháp học máy vào phân tích cảm xúc khách hàng về dòng sản phẩm Herbalife.

[6] Kathuria, P., Sethi, P., & Negi, R. (2022, December). Sentiment analysis on E-commerce reviews and ratings using ML & NLP models to understand consumer behavior. In 2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC) (pp. 1-5). IEEE.

[7] Roland, M. S. A., Simon, J., Misra, S., & Ayeni, F. A Real-Time Sentimental Analysis on E-Commerce Sites in Nigeria Using Machine Learning.

[8] Novitasari, D. (2022). SMEs e-commerce buying intention: how the effect of perceived value, service quality, online customer review, digital marketing and influencer marketing. Journal of Information Systems and Management (JISMA), 1(5), 61-69.

[9] Hiển, B. M., Phát, N. T., Hương, P. T. T., Hương, N. T. B., & Thành, H. T. Phân tích ý kiến khách hàng trực tuyến dựa theo phương pháp học máy.

[10] Moazzam, A., Mushtaq, H., Sarwar, A., Idrees, A., Tabassum, S., & Rehman, K. U. (2021). Customer Opinion Mining by Comments Classification using Machine Learning. International Journal of Advanced Computer Science and Applications, 12(5).

[11] Lin, X. (2020, April). Sentiment analysis of e-commerce customer reviews based on natural language processing. In Proceedings of the 2020 2nd International Conference on Big Data and Artificial Intelligence (pp. 32-36).

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

[13] Shoja, B. M., & Tabrizi, N. (2019). Customer reviews analysis with deep neural networks for e-commerce recommender systems. IEEE access, 7, 119121-119130.

[14] Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE international conference on innovative research and development (ICIRD) (pp. 1-6). IEEE.

[15] Lohse, T., Kemper, J., & Brettel, M. (2017). How online customer reviews affect sales and return behavior–an empirical analysis in fashion e-commerce.

[16] Chauhan, M., & Yadav, D. (2015). Sentimental analysis of product based reviews using machine learning approaches. Journal of Network Communications and Emerging Technologies (JNCET), 5(2), 19-25.

[17] Hoàng, A. C. (2020). Nghiên cứu lý thuyết Naive Bayes và ứng dụng phân loại tài liệu tiếng việt trong thư viện số.

[18] Thủy, N. M., & Tuyền, N. T. M. (2015). Ứng dụng phương pháp phân tích thành phần chính, hồi quy logistic và giản đồ yêu thích trong đánh giá cảm quan sản phẩm sữa gạo. Tạp chí Khoa học Đại học cần Thơ, (37), 11-20.

[19] *API là gì? - Giải thích về Giao diện lập trình ứng dụng - AWS*. (n.d.). AWS. Retrieved December 25, 2023, from https://aws.amazon.com/vi/what-is/api/

[20] *Fleiss Kappa • Simply explained - DATAtab*. (n.d.). DATAtab Statistics Calculator. Retrieved December 25, 2023, from https://datatab.net/tutorial/fleiss-kappa

[21] *Giới thiệu về Support Vector Machine (SVM)*. (n.d.). Viblo. Retrieved December 25, 2023, from https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB

[22] *Machine Learning cơ bản*. (2017, April 9). Machine Learning cơ bản. Retrieved December 25, 2023, from https://machinelearningcoban.com/2017/04/09/smv/#-xay-dung-bai-toan-toi-uu-cho-svm

[23] McCormick, C., Ryan, N., & Gurav, R. (2019, May 14). *BERT Word Embeddings Tutorial · Chris McCormick*. Chris McCormick. Retrieved December 25, 2023, from https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/

[24] Fredriksson, T., Mattos, D. I., Bosch, J., & Olsson, H. H. (2020, November). Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In International Conference on Product-Focused Software Process Improvement (pp. 202-216). Cham: Springer International Publishing.

# PHỤ LỤC

**Bảng tổng hợp những nghiên cứu liên quan**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên bài nghiên cứu** | **Năm xuất bản** | **Tác giả** | **Phương pháp nghiên cứu** | **Điểm mạnh** | **Hạn chế** | **Kết luận** | **Từ khóa** |
| 1 | Sentiment Analysis of Product Reviews for E-Commerce Recommendation based on Machine Learning | 2023 | Manal Loukili, Fayçal Messaoudi, and Mohammed El Ghaz | **- Mô hình:** k-Nearest Neighbors, Random Forest, Catboost, Logistic Regression  **- Dữ liệu:** Bộ dữ liệu liên quan đến cửa hàng thương mại điện tử quần áo phụ nữ | Thực hiện được nhiều mô hình học máy có giám sát nhằm đánh giá được độ chính xác giữa các mô hình và tìm ra được mô hình vượt trội hơn các mô hình còn lại. | Chưa phát hiện được những đánh giá giả mạo làm ảnh hết đến những kết quả mà nhóm phân tích. | Các mô hình đề xuất ở trên đạt độ chính xác từ 80% đến 90%. Mô hình hồi quy logistic vượt trội hơn các mô hình khác về các tham số ma trận nhầm lẫn và điểm AUC. | Classification Algorithms, E-Commerce, Natural Language Processing,  Product Reviews, Sentiment Analysis, Supervised Learning |
| 2 | Sentiment analysis of amazon products using a hybrid Random Forest and Support Vector Machine model. | 2023 | Om Sarulkar, Rahul Pitale, Shivam Tikhe, Sumit Giri, Rohan More | **- Mô hình**: Random forest, SVM  **- Dữ liệu**: tập dữ liệu đánh giá sản phẩm của Amazon | Sử dụng kết hợp Random forest và SVM để tối ưu hoá thuật toán phân tích cảm xúc trong đánh giá sản phẩm của Amazon về đánh giá tiêu cực và tích cực. | Tác giả đi sâu vào phân tích thuật toán mà chưa trình bày rõ ràng vào phân loại và nghiên cứu cảm xúc khách hàng. | Kết quả chỉ ra rằng thuật toán Random forest có độ chính xác 82%, SVM có độ chính xác 87%, nhưng nếu kết hợp cả 2 mô hình thì lên tới 90%. Từ đó đưa ra mô hình tối ưu cho việc đánh giá, phân loại, đưa ra hệ thống đề xuất cho website thương mại điện tử | Sentiment analysis, Sentiment classification, opinion mining, machine learning, polarity classification, supervised algorithms, Amazon classification, Ensemble Learning, Hybrid Machine learning |
| 3 | A Combined Approach of Sentimental Analysis Using Machine Learning Techniques | 2023 | Ketan Gupta, Nasmin Jiwani, Neda Afreen | **- Mô hình:** SVM, Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN), TF-IDF  **- Dữ liệu**: datasets từ Amazon product review, Yelp review, IMDB movie reiview | Bài báo nêu bật lý thuyết cơ bản đằng sau mô hình cũng như các phương pháp tiếp cận mà tác giả đã thực hiện trong nghiên cứu của mình, đưa ra thước đo hiệu suất.  Sử dụng kết hợp mô hình TF-IDF để đánh giá độ quan trọng và loại bỏ các dữ liệu gây nhiễu. | Tác giả chỉ phân tích cảm xúc tiêu cực theo đánh giá 1 và 2 sao, tích cực theo đánh giá 4 và 5 sao mà loại bỏ 3 sao, điều này đặt ra câu hỏi về tính chính xác trong quá trình phân tích | Kết quả chỉ ra rằng thuật toán Random forest có độ chính xác lên tới 78%, vượt trội hơn so với các mô hình còn lại. | sentimental analysis, Amazon, IMDB, yelp, feature extraction |
| 4 | Sentiment Analysis of Women's Clothing Reviews on E-commerce Platforms: A Machine Learning Approach | 2023 | Masfiq Mahmud, Rafit Ahmed Mullick, Muhammad Anas | **- Mô hình:** Random Forest, Support Vector Machine (SVM), Logistic Regression, Naive Bayes Classifier, Gradient Boosting, and Long Short-Term Memory (LSTM)  **- Dữ liệu:** 30.000 điểm dữ liệu được thu thập từ một nền tảng thương mại điện tử ở Bangladesh về lĩnh vực thời trang phụ nữ | Bài nghiên cứu phân tích đánh giá của khách hàng về 3 khía cạnh tích cực, tiêu cực và trung tính; chỉ ra mức độ tương quan giữa thuộc tính sản phẩm và thái độ khách hàng. Tác giả phân loại được những đánh giá bằng tiếng Anh và tiếng Benga. Tác giả cũng đề cập đến việc lọc các đánh giá tiêu cực nhưng được xếp hàng 4-5 sao. Kết quả chỉ ra được sản phẩm nào có đánh giá tích cực cao nhất, sản phẩm có đánh giá tiêu cực nhất. | Bài nghiên cứu chưa mô tả rõ ràng cách tác giả thực hiện các thuật toán mà chỉ đưa ra một bảng kết quả chung. | Tác giả sử dụng 6 thuật toán và kết quả chỉ ra Random Forest có độ chính xác cao nhất, lên đến 96%. Độ chênh lệch giữa thuật toán có độ chính xác cao nhất và độ chính xác thấp nhất là rất lớn, xấp xỉ 20%. | Accuracy, Customer Reviews, E-commerce, Gradient Boosting, LSTM, Machine Learning, Naive Bayes Classifier, Natural Language Processing, Predictive Modeling, Random Forest, Classifier, Sentiment Analysis, Support Vector Machine, Topic Modeling, Women's Clothing |
| 5 | Ứng dụng phương pháp học máy vào phân tích cảm xúc khách hàng về dòng sản phẩm Herbalife | 2023 | Đỗ Thành Đạt | **- Mô hình:** Naive Bayes, Logistic Regression, Support Vector Machine (SVM), XGBoost  **- Phương pháp:** Hold-out  **- Dữ liệu:** 15,483 ý kiến khách hàng về sản phẩm Herbalife trên website amazon.com | Tác giả đã sử dụng 4 thuật toán để nghiên cứu và so sánh kết quả, từ đó chọn ra thuật toán có độ chính xác cao nhất là Logistic Regression. Bài nghiên cứu đã chỉ ra được tỷ lệ đánh giá tích cực và tiêu cực của từng dòng sản phẩm, cho thấy khách hàng quan tâm dòng sản phẩm nào nhất; dòng sản phẩm nào bị đánh giá tiêu cực nhiều nhất. | Bài nghiên cứu chỉ tập trung vào phân tích cảm xúc mang tính tích cực và tiêu cực, chưa quan tâm đến cảm xúc mang tính trung tính. Dữ liệu thu thập chỉ mới tập trung xử lý ngôn ngữ tiếng anh, chưa mở rộng được việc thu thập dữ liệu ngôn ngữ Tiếng Việt. | 4 thuật toán được nghiên cứu đều có độ chính xác rất cao, từ 87%-89% trong đó Logistic Regression có độ chính xác cao nhất. Nghiên cứu này giúp doanh nghiệp có thể ứng dụng xây dựng dashboard theo dõi trải nghiệm khách hàng, hỗ trợ doanh nghiệp ra quyết định kịp thời nhằm cải thiện chất lượng sản phẩm và dịch vụ. | Phân tích cảm xúc khách hàng, xử lý ngôn ngữ tự nhiên, xử lý ngôn ngữ tiếng anh, mô hình học máy, thương mại điện tử |
| 6 | Sentiment analysis on E-commerce reviews and ratings using ML & NLP models to understand consumer behavior | 2022 | Priyanshi Kathuria, Parth Sethi, Rithwick Negi | **- Mô hình:** Logistic Regression, ADA boost, SVM, Naive Bayes và Random Forest  **- Công nghê:** Vader và Text Blob  **- Dữ liệu:** Đánh giá của khách hàng trên các sàn thương mại điện tử Flipkart, Amazon,... về sản phẩm thời trang nữ | Nhóm tác giá thực hiện phân tích với nhiều thông số khác nhau và nhiều mô hình để tìm hiểu liệu khách hàng có thích 1 sản phẩm nào đó hay không. Bên cạnh các mô hình học máy, nhóm tác giả còn sử dụng thêm 2 thư viện của python là Vader và Text Blob để phân tích cụ thể hơn về cảm xúc trong một lượt đánh giá của khách hàng. | Bài báo sử dụng bộ dữ liệu được lấy trên đa dạng các sàn thương mại điện tử nơi có thể có nhiều ngôn ngữ bình luận nhưng tác giả chưa đề cập đến việc xử lý các dữ liệu này, điều này có thể dẫn đến sai số trong quá trình máy học. | Kết quả dự đoán cho thấy với mô hình Logistic Regression cho kết quả có độ chính xác cao nhất ở cả hai nhóm phân tích là bình luận (88,18%) và đánh giá (80,68%). | ML, NLP, Sentiment analysis, Logistic regression, Ada boost, SVM, Naive Bayes, random forest, Vader, Text blob, Consumer behavior |
| 7 | A Real-Time Sentimental Analysis on E-Commerce Sites in Nigeria Using Machine Learning | 2022 | Miriam Shaba, Andeboutom Roland, John Simon, Sanjay Misra, Foluso Ayeni | **- Mô hình:** Naive bayes, SVM, Logistic regression  **- Dữ liệu:** 54135 dataset lấy từ 3 website Jumia, Konga, Jiji | Sử dụng dữ liệu dựa trên thời gian thực, có nhiều hình ảnh trực quan được đưa ra.  Trong các paper khác, giai đoạn data preprocessing thu thập các nhận xét hoặc đánh giá dữ liệu bằng cách xóa các ký hiệu, dấu câu, URL,.. Nhưng trong paper này, tác giả giữ nguyên dấu câu và các ký hiệu khác tương quan trong đánh giá tiêu cực hoặc tích cực do việc đặt dấu câu vào bất kỳ phần nào của nhận xét có thể có ý nghĩa khác ngay cả khi các từ giống nhau. | Tác giả chỉ đưa ra biểu đồ mà chưa đưa ra con số cụ thể trong đánh giá comment tích cực và tiêu cực. Các thuật toán được đưa ra chưa được giải thích sâu và cụ thể về cách triển khai, tác giả chủ yếu chỉ đi sâu và việc phân tích quá trình thực hiện | Kết quả cho thấy độ chính xác của các thuật toán được đưa ra lên tới 90% trong việc xác định trang web, sản phẩm tốt hơn . Từ đó người dùng có thể tìm ra được sản phẩm, hoặc nhãn hiệu sản phẩm tốt nhất để mua và trang web thương mại điện tử tốt nhất để mua hàng. | Sentimental analysis, machine learning, e-commerce |
| 8 | SMEs E-commerce Buying Intention:How the Effect of Perceived Value, Service Quality, Online Customer Review , Digital Marketing and Influencer Marketing | 2022 | Dewiana Novitasari | - **Mô hình:** Structural Equation Modeling (SEM) (Mô hình cấu trúc tuyến tính) với công cụ SmartPLS 4.0  **- Dữ liệu:** 550 nhân viên của các công ty vừa và nhỏ (SMEs) tại Banten (Indonesia) | Không chỉ xem xét 1 yếu tố OCRs mà còn nghiên cứu chung hơn về ảnh hưởng của Digital Marketing, Influencer Marketing, Service Quality, Perceived Value đến ý định mua hàng của khách hàng. | Bài báo chỉ khẳng định và chứng minh lại về việc online customer reviews sẽ ảnh hưởng đến mua sắm online chứ chưa đưa ra những khám phá mới mẻ hơn | Digital Marketing, influencer marketing, online customer reviews có ảnh hưởng tới ý định mua hàng của khách hàng. Những khách hàng có ý định mua sản phẩm sẽ tham khảo ý kiến review và phản hồi từ những người đã mua trước đó. | SMEsDigital Marketing, Online Buying intention, Online Customer Review |
| 9 | Phân tích ý kiến khách hàng trực tuyến dựa vào phương pháp học máy | 2021 | Hồ Trung Thành,  Bùi Minh Hiển,  Phạm Thị Thiên Hương,  Nguyễn Thành Phát,  Nguyễn Thị Bảo Hương | - **Phương pháp:** học máy có giám sát  **- Mô hình:** Hồi quy Logictic, Random Forest  **- Dữ liệu:** từ Lazada | Nhóm tác giả thực hiện được 2 mô hình để so sánh và đánh giá độ chính xác. | Chưa phân tích dữ liệu trong thời gian thực, thời gian huấn luyện và thời gian dự đoán để nhà quản trị thương mại điện tử đưa ra quyết định nhanh chóng. | Quá trình phân tích, kết quả, mô hình đề xuất có thể nhận biết được cảm xúc tích cực và tiêu cực trong ý kiên khách hàng đã ảnh hưởng 90% đên ứng dụng Lazada, điển hình tập trung vào các tính năng, chất lượng cùa ứng dụng. | Phân tích ý kiến khách hàng, thương mại điện tử, khách hàng trực tuyến, phân tích cảm xúc, học máy có giám sát. |
| 10 | Customer Opinion Mining by Comments Classification using Machine Learning | 2021 | Moazzam Ali, Farwa yasmine, Husnain Mushtaq, Abdullah Sarwar, Adil Idrees, Sehrish Tabassum, Dr.BaburHayyat, Khalil Ur Rehman | **- Mô hình:** Bow and N-Gram, Association rule mining, Logistic Regression, K-nearest neighbor (KNN), Neural Networks, Descision Trees  **- Dữ liệu:** 2400 ý kiến | Tác giả sử dụng mô hình N-Gram kết hợp cùng các thuật toán SVM, KNN và Naive Bayes tăng độ chính xác phân loại ý kiến của khách hàng so với chỉ sử dụng một thuật toán đơn lẻ. Tác giả cũng đã nghiên cứu ý kiến của khách hàng trong 4 khía cạnh convenience, variety, time and trust. | Nghiên cứu này không tập trung vào một ngành hàng cố định. | Trong nghiên cứu này, SVM cùng với kỹ thuật BOW và bi-gram có độ chính xác cao nhất. Điều đó giúp khách hàng và người bán tiếp cận nhanh chóng đến ý kiến của những khách hàng khác trong 4 khía cạnh là convenience, variety, time and trust. | Customer comments, behavior mining, data mining, machine learning |
| 11 | Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing | 2020 | Xiaoxin Lin | **- Mô hình**: Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost và LightGBM  **- Dữ liệu:** Bộ dữ liệu Bình luận của khách hàng mua quần áo nữ trên sản thương mại điện tử | Tác giả thực hiện nghiên cứu sử dụng 5 mô hình học máy để đưa ra các kết quả và so sánh một cách khách quan nhất. Tác giả đã thực hiện tính độ tương quan của bình luận của khách hàng và sản phẩm đề xuất để xem liệu một sản phẩm được hệ thống đề xuất có được người dùng yêu thích hay không. | Tác giả thực hiện nghiên cứu sử dụng 5 mô hình học máy để đưa ra các kết quả và so sánh một cách khách quan nhất. Tác giả đã thực hiện tính độ tương quan của bình luận của khách hàng và sản phẩm đề xuất để xem liệu một sản phẩm được hệ thống đề xuất có được người dùng yêu thích hay không. | LightGBM đua ra kết quả cao nhất với độ chính xác là 98% và giá trị AUC là 96%. Các mô hình Ridge Regression, Linear Kernel SVM và XGboost cũng cho ra kết quả tương tự. Dù RBF Kernel SVM cho kết quả thấp nhất nhưng độ chính xác vẫn lên đến 81% | E-commerce, Reviews, Machine learning, Natural language processing |
| 12 | Sentiment Analysis using Ramdom Forest Algorithmonline Social media based | 2019 | Bahrawi | **- Mô hình:** Random forest  **- Dữ liệu:** Dữ liệu hàng không từ Kaggle |  |  | Bài nghiên cứu giải quyết một vấn đề cơ bản của phân tích tình cảm - phân loại tình cảm phân cực đạt được độ chính xác khoảng 75%. | sentiment analysis; random forest algorithm; classification; machine learnings. |
| 13 | Customer Reviews Analysis With Deep  Neural Networks for E-Commerce  Recommender Systems | 2019 | Babak Maleki Shoja, Nasseh Tabrizi | **- Mô hình:** Latent Dirichlet Allocation (LDA), Deep neural networks, Collaborative Filtering  **- Dữ liệu:** Tập dữ liệu từ Amazon | So sánh hiệu suất của mô hình này với ba mô hình cơ sở khác, bao gồm MF, Chủ đề và yếu tố ẩn (HTF) và Đánh giá cuộc họp xếp hạng (RMR). |  |  | Mô hình này hoạt động tốt hơn các mô hình hiện đại này đối với hầu hết các bộ dữ liệu. |
| 14 | Sentiment Analysis on Large Scale Amazon Product Reviews | 2018 | Tanjim Ul Haque, Nudrat Nawal Saber, Faisal Muhammad Shah | **- Mô hình:** Linear support Vector machine, Multinomial Naïve Bayes, Stochastic Gradient Descent, Random Forest, Logistic regression và Decision tree  **- Dữ liệu:** Bộ dữ liệu đánh giá của khách hàng trên các sản phẩm điện tử, điện thoại và phụ kiện và nhạc cụ từ Amazon | Ngoài mô hình học máy có giám sát, nhóm tác giả còn sử dụng kết hợp 2 phương thức trích xuất đặc trưng. | Các bình luận được thu thập do có sự hạn chế từ Amazon nên không có đủ số lượng để xem là những bình luận công khai dựa trên thực tế. Ngoài ra, việc gắn nhãn các bình luận bỏ qua label Neutral làm cho mất sự cân bằng trong các nhãn dán và có thể đưa ra kết quả chênh lệch | Nhóm tác giả tiến hành phân tích dự đoán trên 3 nhóm sản phẩm riêng biệt và đều đưa ra kết quả có độ chính xác cao nhất thông qua mô hình Linear support Vector Machine, lần lượt là Điện thoại và phụ kiện (93,57%), Nhạc cụ (94,02%) và Đồ điện tử (93,52%) | Sentiment analysis, pool based active learning, feature extraction, text classification, machine learning |
| 15 | How online customer review affect sales and return behavior - An empirical Analysis in Fashion E-Commerce | 2017 | Tobias Lohse, Jan Kemper, Malte Brettel | **- Mô hình:** 2SLS log-log-regression model  **- Dữ liệu:** a leading fashion e-commerce company selling shoes- and apparel products across the European market | Tập hợp dữ liệu giao dịch thực tế lớn: dữ liệu giao dịch đa dạng (2,5 tỷ lượt nhấp vào trang, 46 nghìn sản phẩm khác nhau, 700 nhãn hiệu, 40 danh mục sản phẩm, 72 triệu mặt hàng đã bán và 33 triệu mặt hàng bị trả lại) với bộ OCR lớn (0,9 triệu) . | Mặc dù tập data lớn nhưng mà chỉ hạn chế ở 1 company ở Europe. Và khi hạn chế ở châu Âu thì văn hóa, lối sống cũng rất khác, nên mở rộng nghiên cứu ra thêm | OCR dương có thể dẫn đến tỷ lệ hoàn hàng thấp hơn, doanh thu cao hơn và tỷ lệ chuyển đổi tốt hơn. Tiết lộ tác động yếu hơn của OCR trên thiết bị di động so với kênh bán hàng trên máy tính để bàn. Nhìn thấy tác động đáng kể của sự tham gia của sản phẩm trong bối cảnh này chẳng hạn như ảnh hưởng của OCR tích cực đối với các sản phẩm có mức độ tham gia cao sẽ mạnh hơn so với ngược lại. Hơn nữa, nhận thấy sự ủng hộ cho các tuyên bố từ tài liệu về tín hiệu thương hiệu, rằng OCR quan trọng hơn đối với các thương hiệu yếu hơn là đối với các thương hiệu mạnh. | eWOM, Online Customer Reviews, Product Returns, Mobile Shopping, Product Involve-ment, Brand Equity |
| 16 | Sentimental analysis of product based reviews using Machine Learning approaches | 2015 | Manvee Chauhan, Divakar Yadav | **- Mô hình**: Naive Bayes, SVM  **- Dữ liệu:** tập dữ liệu về các sản phẩm như điện thoại, ipod trên các web .cnet.com, downlaod.com, reviewcentre.com, zdnet.com, epinions.com và Consumerreview.com | Bài nghiên cứu sử dụng Naive Bayes và SVM để phân tích. Tác giả đã phân tích được cảm xúc mang tính tích cực, tiêu cực và cả trung tính | Tác giả chỉ dùng 2 thuật toán để phân tích nên chưa có tính thuyết phục.  Chỉ phân tích đánh giá bằng tiếng anh, chưa thực hiện được trên nhiều ngôn ngữ. | Trong 2 thuật toán tác giả nghiên cứu, Naive Bayes có độ chính xác 84,02% cao hơn SVM với 80,2%. Mô hình nghiên cứu này rất hữu ích để xác định và dự đoán xu hướng hiện tại và tương lai. | Machine learning, Semantic Orientation, Sentiment Analysis, Support Vector Machine, Naive Bayes |