

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**



VÕ THỊ NGỌC CHÂU

**XÂY DỰNG MÔ HÌNH DỰ ĐOÁN CẢM XÚC KHÁCH
HÀNG**

**KHÓA LUẬN TỐT NGHIỆP
NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ**

TP. HỒ CHÍ MINH, 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH



VÕ THỊ NGỌC CHÂU

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN CẢM XÚC KHÁCH
HÀNG

Mã số sinh viên: 2154053001

KHÓA LUẬN TỐT NGHIỆP
NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ

Giảng viên hướng dẫn: ThS. Hồ Hương Thiên

TP. HỒ CHÍ MINH, 2025

TRƯỜNG ĐẠI HỌC MỞ CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
THÀNH PHỐ HỒ CHÍ MINH Độc lập – Tự do – Hạnh phúc
KHOA CÔNG NGHỆ THÔNG TIN

GIẤY XÁC NHẬN

Tôi tên là: Võ Thị Ngọc Châu.....

Ngày sinh: 24/02/2003..... Nơi sinh: An Giang.....

Chuyên ngành: HTTTQL..... Mã sinh viên: 2154053001.....

Tôi đồng ý cung cấp toàn văn thông tin đồ án/ khóa luận tốt nghiệp hợp lệ về bản quyền cho Thư viện Trường Đại học Mở Thành phố Hồ Chí Minh. Thư viện Trường Đại học Mở Thành phố Hồ Chí Minh sẽ kết nối toàn văn thông tin đồ án/ khóa luận tốt nghiệp vào hệ thống thông tin khoa học của Sở Khoa học và Công nghệ Thành phố Hồ Chí Minh.

Ký tên
(Ghi rõ họ và tên)

Võ Thị Ngọc Châu

**Ý KIẾN CHO PHÉP BẢO VỆ ĐỒ ÁN/ KHÓA LUẬN TỐT NGHIỆP
CỦA GIẢNG VIÊN HƯỚNG DẪN**

Giảng viên hướng dẫn:

Sinh viên thực hiện: **Lớp:**

Ngày sinh: **Nơi sinh:**

Tên đề tài:

.....

.....

.....

.....

**Ý kiến của giảng viên hướng dẫn về việc cho phép sinh viên được bảo vệ đồ án/
khóa luận trước Hội đồng:**

.....

.....

.....

.....

.....

.....

.....

.....

.....

Thành phố Hồ Chí Minh, ngày ... tháng ... năm

Người nhận xét

LỜI CẢM ƠN

Em xin trân trọng gửi lời cảm ơn đến quý thầy cô Trường Đại học Mở TP.HCM, những người đã luôn tận tâm dìu dắt và truyền đạt cho em những kiến thức quý báu trong suốt quá trình học tập. Nhờ có nền tảng kiến thức vững chắc ấy, em đã có thể tự tin áp dụng vào khóa luận tốt nghiệp của mình.

Bài báo cáo này không chỉ là kết quả của riêng em mà còn là sự kết tinh từ sự hỗ trợ tận tình của nhà trường, sự hướng dẫn chu đáo của quý thầy cô cũng như sự giúp đỡ quý báu từ các bạn/anh/chị để em có thể tiến bộ được như ngày hôm nay.

Em cũng xin chân thành cảm ơn thầy ThS. Hồ Hương Thiên, sự hướng dẫn, hỗ trợ nhiệt tình từ thầy đã giúp đỡ em rất nhiều trong quá trình hoàn thành khóa luận tốt nghiệp.

Với thái độ nghiêm túc, dù đã hoàn thành báo cáo với sự nỗ lực cao nhất nhưng sẽ không thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bờ ngờ. Em rất mong nhận được sự đóng góp ý kiến của quý thầy cô để bài báo cáo được hoàn thiện hơn.

Em xin chân thành cảm ơn!

TP.HCM, Ngày 04 tháng 05 năm 2025

Võ Thị Ngọc Châu

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TÓM TẮT KHÓA LUẬN

Nhằm hỗ trợ doanh nghiệp và người dùng hiểu rõ hơn về phản hồi của khách hàng trên các nền tảng thương mại điện tử, đề tài này tập trung xây dựng mô hình phân tích cảm xúc trong bình luận sản phẩm. Sau đó triển khai một hệ thống đơn giản dưới dạng một trang web trực tuyến, cho phép người dùng nhập các bình luận và nhận được kết quả phân tích cảm xúc tương ứng (tích cực hoặc tiêu cực). Hệ thống này được phát triển bằng ngôn ngữ lập trình Python, sử dụng framework Django cho backend, cùng các công nghệ web như HTML, CSS và JavaScript cho phần giao diện.

Mục tiêu chính của đề tài là phát triển mô hình học máy có khả năng phân loại cảm xúc trong bình luận người dùng một cách hiệu quả. Đề tài đã áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP), các thuật toán học máy (Machine Learning) và học sâu (Deep Learning), trong đó mô hình LSTM cho thấy hiệu quả vượt trội so với các phương pháp truyền thống. Dữ liệu huấn luyện được thu thập từ các bình luận thực tế trên sàn thương mại điện tử Shopee, đảm bảo phản ánh đúng ngôn ngữ tự nhiên và đa dạng trong cách thể hiện cảm xúc của người dùng.

Kết quả thực nghiệm cho thấy mô hình đạt độ chính xác cao trên tập dữ liệu thử nghiệm, đặc biệt là khi sử dụng LSTM với các tham số tối ưu. Hệ thống không chỉ hỗ trợ doanh nghiệp trong việc theo dõi và phân tích mức độ hài lòng của khách hàng, mà còn góp phần nâng cao chất lượng sản phẩm, dịch vụ và trải nghiệm người dùng. Qua đó, nghiên cứu đóng vai trò quan trọng trong việc khai thác hiệu quả dữ liệu phi cấu trúc, đồng thời tạo nền tảng cho các ứng dụng phân tích phản hồi và cải thiện chiến lược kinh doanh dựa trên dữ liệu.

ABSTRACT

To support businesses and users in better understanding customer feedback on e-commerce platforms, this study focuses on building a sentiment analysis model for product reviews. A simple web-based system is then deployed, allowing users to input comments and receive corresponding sentiment predictions (positive or negative). The system is developed using the Python programming language, with Django as the backend framework, and HTML, CSS, and JavaScript for the frontend interface.

The main objective of this project is to develop a machine learning model capable of effectively classifying user sentiment in product reviews. The study applies natural language processing (NLP) techniques along with machine learning and deep learning algorithms, with the LSTM model demonstrating superior performance compared to traditional approaches. The training data is collected from real customer reviews on the Shopee e-commerce platform, reflecting the natural and diverse ways in which users express their opinions.

Experimental results show that the model achieves high accuracy on the test dataset, especially when using LSTM with optimized parameters. The system not only assists businesses in monitoring and analyzing customer satisfaction but also contributes to improving product quality, services, and overall user experience. As such, this research plays an important role in leveraging unstructured data effectively and lays the foundation for feedback analysis applications and data-driven business strategy optimization.

MỤC LỤC

TÓM TẮT KHÓA LUẬN	5
ABSTRACT	6
DANH MỤC TỪ VIẾT TẮT	10
DANH MỤC HÌNH VẼ	11
DANH MỤC BẢNG	12
MỞ ĐẦU	13
Chương 1. TỔNG QUAN VỀ ĐỀ TÀI	14
1.1. Lý do chọn đề tài	14
1.2. Mục tiêu của đề tài	15
1.3. Đối tượng nghiên cứu của đề tài	15
1.4. Phương pháp nghiên cứu	16
1.5. Bố cục của báo cáo	16
Chương 2. CƠ SỞ LÝ THUYẾT	18
2.1. Phân tích cảm xúc	18
2.1.1. Tổng quan về phân tích cảm xúc	18
2.1.2. Phân tích cảm xúc ứng dụng học máy	18
2.2. Học máy	19
2.2.1. Giới thiệu	19
2.2.2. Quy trình hoạt động	19
2.2.3. Phân loại	21
2.2.4. Các thuật toán học máy	22
2.3. Học sâu	28
2.3.1. Tổng quan về học sâu	28

2.3.2. Kiến trúc mạng nơ-ron nhân tạo	28
2.3.3. Hàm kích hoạt	28
2.3.4. Kiến trúc học sâu trong NLP	29
2.4. Các phương pháp trích xuất đặc trưng	31
2.4.1. Bag-of-words	31
2.4.2. TF-IDF	32
2.4.3. N-grams	33
2.5. Phương pháp đánh giá mô hình	34
2.5.1. Accuracy	34
2.5.2. ROC-AUC	35
2.6. Django Framework:	37
2.6.1. Tổng quan về Django	37
2.6.2. Kiến trúc MTV (Model - Template - View)	37
2.6.3. Các thành phần chính trong Django	37
2.6.4. Ưu điểm của Django	38
Chương 3. HỆ THỐNG DỰ ĐOÁN CẢM XÚC	39
3.1. Giới thiệu hệ thống	39
3.2. Kiến trúc hệ thống	39
3.2.1. Thu thập dữ liệu	39
3.2.2. Đào tạo mô hình	40
3.2.3. Mô hình dự đoán:	60
3.3. Kết quả đạt được	60
3.3.1. Chức năng	60
3.3.2. Kết quả mô hình	61
Chương 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	63

4.1. Kết luận	63
4.2. Hướng phát triển	63
TÀI LIỆU THAM KHẢO	65
PHỤ LỤC	67

DANH MỤC TỪ VIẾT TẮT

Ký hiệu viết tắt	Chữ viết đầy đủ
API	Application Programming Interface
BoW	Bag-of-Words
GMV	Gross Merchandise Value
K-NN	K-Nearest Neighbor
LR	Logistic Regression
NLP	Natural Language Processing
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
TMĐT	Thương mại điện tử
URL	Uniform Resource Locator

DANH MỤC HÌNH VẼ

Hình 1 . Quy trình hoạt động của học máy (Nguồn: internet)	21
Hình 2 . Thuật toán Logistic Regression (Trích nguồn Internet)	23
Hình 3 . Thuật toán SVM (Trích nguồn Internet)	24
Hình 4 . Thuật toán Random Forest (Trích nguồn Internet)	25
Hình 5 . Ma trận từ theo mô hình Bag-of-Words	32
Hình 6 . Danh mục dữ liệu được thu thập	39
Hình 7 . Minh họa chuẩn mã hóa dữ liệu	41
Hình 8 . Đánh giá chưa link url.	41
Hình 9 . Các emojis trong đánh giá.	42
Hình 10 . Xử lý các từ bị lặp ký tự cuối.	43
Hình 11 . Đánh giá vô nghĩa.	44
Hình 12 . Các từ viết tắt.	44
Hình 13 . Một số hư từ.	45
Hình 14 . Xử lý hư từ.	45
Hình 15 . Độ dài của đánh giá.	46
Hình 16 . Tách từ.	50
Hình 18 . Chuẩn bị dữ liệu cho Emoji sentiment model.	52
Hình 20 . Train cross-validation và lựa chọn hyperparameter set cho từng model.	56
Hình 21 . Ma trận nhầm lẫn của mô hình LSTM lần 1.	58
Hình 22 . Trang chủ hệ thống dự đoán cảm xúc khách hàng.	61
Hình 23 . Hệ thống trả kết quả với đánh giá 5 sao.	61
Hình 24 . Hệ thống trả kết quả với đánh giá 1 sao.	61

DANH MỤC BẢNG

Bảng 1 . Các thành phần chính trong Django.	38
Bảng 2 . Mô tả các thuộc tính của bộ dữ liệu.	40
Bảng 3 . Đánh giá có độ dài > 400.	50
Bảng 4 . Số lượng bình luận ở mỗi rating.	51
Bảng 5 . Số lượng bình luận ở mỗi nhãn	51
Bảng 6 . Số đánh giá sau cân bằng ở mỗi nhãn	51
Bảng 7 . Kết quả so sánh các mô hình dựa vào độ chính xác trên tập huấn luyện.	53
Bảng 8 . Kết quả so sánh các mô hình dựa vào độ chính xác trên tập kiểm tra.	53
Bảng 9 . Đánh giá các mô hình Emoji sentiment comment.	54
Bảng 10 . Kết quả so sánh top 6 mô hình dựa vào độ chính xác trên tập huấn luyện ...	55
Bảng 11 . Kết quả so sánh top 6 mô hình dựa vào độ chính xác trên tập kiểm tra.	56
Bảng 13 . Kết quả huấn luyện mô hình LSTM lần 1.	58
Bảng 14 . Độ chính xác của các mô hình kết hợp.	59

MỞ ĐẦU

Trong bối cảnh xã hội hiện đại không ngừng phát triển, đặc biệt là trong các lĩnh vực kinh tế và công nghệ, nhu cầu phân tích cảm xúc trong văn bản ngày càng trở nên thiết yếu. Lĩnh vực này hiện đang được ứng dụng rộng rãi trong nhiều hoạt động thực tiễn như quản lý quan hệ khách hàng, xây dựng và bảo vệ thương hiệu, khảo sát dư luận xã hội cũng như đánh giá phản hồi của người tiêu dùng đối với sản phẩm, dịch vụ. Trong môi trường kinh doanh cạnh tranh ngày càng gay gắt, việc nắm bắt cảm xúc và ý kiến của người dùng đã trở thành một yếu tố then chốt giúp các doanh nghiệp thích ứng nhanh chóng với thị trường và đưa ra các chiến lược phù hợp.

Sự quan tâm của cộng đồng đối với các sản phẩm như phim ảnh, sách, thiết bị điện tử và nhiều mặt hàng tiêu dùng khác có ảnh hưởng lớn đến thành công của các doanh nghiệp. Do đó, cả giới học thuật và doanh nghiệp đều chú trọng đến việc xây dựng các hệ thống phân tích cảm xúc tự động nhằm hiểu rõ hơn nhu cầu, thị hiếu và mức độ hài lòng của khách hàng. Các hệ thống này không chỉ hỗ trợ công tác tiếp thị và cải thiện sản phẩm, mà còn góp phần nâng cao trải nghiệm người dùng thông qua việc tối ưu hóa quá trình ra quyết định dựa trên dữ liệu.

Với mục tiêu đó, đề tài này được thực hiện nhằm khai thác khả năng của các thuật toán Học máy trong việc phân tích dữ liệu đánh giá, phản hồi của khách hàng. Thông qua đó, hệ thống có thể nhận diện và phân loại cảm xúc, giúp doanh nghiệp hiểu rõ hơn tâm lý người tiêu dùng, đưa ra các chiến lược kinh doanh hiệu quả hơn, đồng thời hỗ trợ cải tiến sản phẩm và hoạt động truyền thông một cách tối ưu.

Chương 1. TỔNG QUAN VỀ ĐỀ TÀI

1.1. Lý do chọn đề tài

Trong bối cảnh xã hội ngày càng được thay thế và hỗ trợ bởi các công nghệ tiên tiến, nhu cầu và hành vi tiêu dùng của con người cũng liên tục thay đổi và nâng cao. Đại dịch Covid-19 không chỉ tác động mạnh mẽ đến đời sống sinh hoạt hàng ngày mà còn khiến nhiều người thay đổi thói quen, trong đó đáng chú ý nhất là xu hướng chuyển dịch sang mua sắm trực tuyến. Theo báo cáo của Cục Thương mại điện tử và Kinh tế số (iDEA), thị trường thương mại điện tử (TMĐT) tại Việt Nam đang phát triển nhanh chóng với tốc độ tăng trưởng từ 25–30% mỗi năm.

Sự chuyển biến này mở ra cơ hội lớn cho các doanh nghiệp, đặc biệt là các thương hiệu trong lĩnh vực thời trang, mở rộng hoạt động kinh doanh trực tuyến nhằm tiếp cận khách hàng hiệu quả hơn. Với sự đa dạng ngày càng cao của sản phẩm quần áo, người tiêu dùng có thể dễ dàng lựa chọn các mặt hàng phù hợp với nhu cầu, công việc và sở thích cá nhân. Việc mua sắm trở nên thuận tiện hơn khi người dùng chỉ cần vài thao tác trên điện thoại thông minh thay vì phải đến trực tiếp các cửa hàng truyền thống. Chính sự tiện lợi này đã trở thành động lực thúc đẩy ngành hàng thời trang trên các sàn TMĐT phát triển mạnh mẽ.

Số liệu từ YouNet ECI cho thấy tổng giá trị giao dịch (GMV) của bốn sàn TMĐT lớn (Shopee, TikTok Shop, Lazada, Tiki) đạt 87,37 nghìn tỷ đồng, tăng 10,4% so với quý trước. Đáng chú ý, ngành hàng "Thời trang & Phụ kiện" chiếm tỷ trọng lớn nhất với GMV đạt 22,679 nghìn tỷ đồng, cho thấy tiềm năng phát triển vượt bậc, đặc biệt trên nền tảng Shopee – nơi có sự kết nối mạnh mẽ với hàng loạt nhà cung cấp trong và ngoài nước. Tuy nhiên, trước sự cạnh tranh khốc liệt và nhu cầu ngày càng khắt khe của người tiêu dùng, các doanh nghiệp buộc phải hiểu rõ hơn về mức độ hài lòng cũng như sở thích cá nhân của khách hàng để duy trì lợi thế cạnh tranh.

Từ thực tế đó, đề tài **“Xây dựng mô hình dự đoán cảm xúc khách hàng”** đã được thực hiện với mong muốn áp dụng các kỹ thuật học máy (Machine Learning) để khai thác thông tin cảm xúc từ những phản hồi, bình luận của khách hàng. Thông qua dự án

này, đề tài kỳ vọng có thể góp phần nâng cao trải nghiệm mua sắm trực tuyến, đồng thời hỗ trợ doanh nghiệp trong việc đánh giá độ tin cậy của sản phẩm, điều chỉnh chiến lược kinh doanh và tối ưu hóa chất lượng dịch vụ dựa trên cảm nhận thực tế từ người tiêu dùng.

1.2. Mục tiêu của đề tài

Đề tài hướng đến việc nghiên cứu và phát triển một mô hình học máy (machine learning) có khả năng phân loại cảm xúc trong bình luận khách hàng với độ chính xác cao, tập trung vào hai nhóm chính là tích cực và tiêu cực.

Để thực hiện điều này, đề tài ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) nhằm làm sạch dữ liệu văn bản. Từ đó, trích xuất được những thông tin có lợi nhằm hỗ trợ hiệu quả cho quá trình thấu hiểu khách hàng của doanh nghiệp

Quá trình thấu hiểu khách hàng được thực hiện thông qua việc trực quan hóa dữ liệu. Với các biểu đồ, báo cáo, “cảm xúc” của khách hàng sẽ được tiếp cận dễ dàng hơn. Từ đó, cung cấp thông tin chi tiết, hỗ trợ doanh nghiệp đưa ra quyết định cải thiện dịch vụ và gia tăng sự hài lòng của khách hàng..

Xây dựng mô hình với nhiều thuật toán phù hợp, tiến hành thực nghiệm, đánh giá độ chính xác của các mô hình và so sánh hiệu quả giữa các mô hình khác nhau để tìm ra mô hình tốt nhất.

1.3. Đối tượng nghiên cứu của đề tài

Đối tượng nghiên cứu của đề tài bao gồm:

Đánh giá của khách hàng trên sàn TMĐT Shopee: Nghiên cứu các đánh giá của khách hàng trên sàn TMĐT Shopee về ngành hàng Thời trang (thời trang nam và thời trang nữ). Đối tượng này là nguồn dữ liệu chính để phân tích, trực quan và xây dựng mô hình dự đoán cảm xúc khách hàng.

Kỹ thuật xử lý NLP và thuật toán học máy: Tập trung vào các kỹ thuật xử lý NLP và các thuật toán học máy để phân loại cảm xúc dựa trên các đánh giá.

Phương pháp đánh giá mô hình: Nghiên cứu các phương pháp và chỉ số đánh giá hiệu suất của mô hình phân loại, như độ chính xác, độ nhạy, và độ đặc hiệu, nhằm đảm bảo rằng mô hình hoạt động hiệu quả và đáng tin.

1.4. Phương pháp nghiên cứu

Phương pháp thống kê và so sánh: Thu thập và phân tích dữ liệu về lượt đánh giá (rating) và bình luận từ các sản phẩm thời trang trên Shopee. Tiến hành so sánh mức độ đánh giá giữa các sản phẩm, đồng thời gán nhãn cho các bình luận nhằm phân loại thành phản hồi tích cực hoặc tiêu cực.

Phương pháp phân tích và tổng hợp: Sau quá trình thống kê, dữ liệu được phân tích chi tiết để nhận diện xu hướng và đặc điểm chung. Tiếp theo, tổng hợp thông tin từ bảng phân loại nhằm rút ra những kết luận quan trọng về cảm xúc khách hàng đối với sản phẩm.

Nghiên cứu lý thuyết về xử lý ngôn ngữ tự nhiên, các phương pháp trích xuất đặc trưng đối với dữ liệu văn bản, nghiên cứu lý thuyết về các bộ phân lớp như Logistic Regression, SVM, Random Forest, XGBoost...

Nghiên cứu thực nghiệm: áp dụng các bộ phân lớp để xây dựng mô hình dự đoán.

1.5. Bố cục của báo cáo

Bám sát mục tiêu, đối tượng, phạm vi nghiên cứu đã xác định, bố cục của đề tài được thiết kế thành 4 chương sau:

Chương 1 - Tổng quan về đề tài: Giới thiệu tổng quan về đề tài, lý do chọn đề tài mục tiêu nghiên cứu, phương pháp nghiên cứu, phạm vi và đối tượng nghiên cứu.

Chương 2 - Cơ sở lý thuyết: Trình bày các cơ sở lý thuyết, khái niệm cơ bản được sử dụng trong đề tài về phân tích cảm xúc, xử lý ngôn ngữ tự nhiên, các thuật toán trích xuất đặc trưng và các mô hình học máy, học sâu.

Chương 3 - Hệ thống dự đoán cảm xúc khách hàng: Tập trung mô tả chi tiết về hệ thống dự đoán cảm xúc khách hàng.

Chương 4 - Kết luận và hướng phát triển: Tổng hợp kết quả của quá trình nghiên cứu khi so sánh các mô hình và huấn luyện mô hình, hạn chế và hướng phát triển thêm sau này.

Chương 2. CƠ SỞ LÝ THUYẾT

2.1. Phân tích cảm xúc

2.1.1. Tổng quan về phân tích cảm xúc

Phân tích cảm xúc (Sentiment Analysis) là một lĩnh vực thuộc Xử lý ngôn ngữ tự nhiên (NLP), tập trung vào việc xác định, trích xuất và phân loại cảm xúc, quan điểm, thái độ hoặc ý kiến trong văn bản.

Phân tích cảm xúc là quá trình sử dụng các phương pháp tính toán để xác định và đánh giá ý kiến, cảm xúc cũng như thái độ thể hiện trong văn bản. Nói cách khác, đây là quá trình phân loại nội dung văn bản theo các nhãn cảm xúc như tích cực, tiêu cực hoặc trung tính.

Phân tích cảm xúc có nhiều ứng dụng thực tiễn, đặc biệt là trong thương mại điện tử và nghiên cứu thị trường. Nó giúp doanh nghiệp và tổ chức nắm bắt phản hồi của khách hàng về sản phẩm, cải thiện dịch vụ, thương hiệu, xây dựng chiến lược tiếp thị hiệu quả và hỗ trợ ra quyết định dựa trên dữ liệu. Một đánh giá tích cực có thể giúp thúc đẩy doanh số bán hàng, trong khi những đánh giá tiêu cực có thể cung cấp thông tin quan trọng về những điểm cần cải thiện.

2.1.2. Phân tích cảm xúc ứng dụng học máy

Việc phân tích cảm xúc thủ công trở nên không khả thi khi số lượng đánh giá trên các nền tảng thương mại điện tử tăng nhanh. Do đó, các phương pháp học máy (Machine Learning) đã được phát triển để tự động hóa quá trình này, giúp doanh nghiệp phân tích hàng triệu đánh giá một cách nhanh chóng và chính xác.

Học máy là một lĩnh vực trong trí tuệ nhân tạo, nhằm giúp hệ thống tự động phân tích dữ liệu mà không cần tới sự can thiệp trực tiếp từ con người thông qua việc lập trình, giúp hệ thống nhận dạng dữ liệu và đánh giá chúng, kết hợp với tương tác với dữ liệu đã được đào tạo để đưa ra các dự đoán.

2.2. Học máy

2.2.1. Giới thiệu

Học máy (Machine Learning - ML) là một lĩnh vực cốt lõi trong trí tuệ nhân tạo (Artificial Intelligence - AI), tập trung vào việc phát triển các hệ thống có khả năng tự động học hỏi từ dữ liệu và cải thiện hiệu suất theo thời gian mà không cần được lập trình tường minh cho từng nhiệm vụ cụ thể. Thay vì viết các quy tắc cố định, các nhà phát triển xây dựng mô hình có thể suy luận và dự đoán dựa trên các mẫu (patterns) đã học được từ dữ liệu lịch sử.

Mục tiêu chính của học máy là xây dựng các mô hình toán học có khả năng tổng quát hóa tốt, tức là đưa ra dự đoán chính xác trên các dữ liệu chưa từng thấy. Các ứng dụng điển hình của học máy có thể kể đến như: phân loại email rác, nhận diện khuôn mặt, dự báo thời tiết, phân tích cảm xúc trong văn bản, và nhiều hơn nữa.

Với khả năng tự động hóa và thích ứng với dữ liệu, học máy đã trở thành nền tảng cho nhiều công nghệ hiện đại, đặc biệt là trong các bài toán xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), nơi mà dữ liệu đầu vào mang tính phi cấu trúc và giàu ngữ nghĩa như văn bản, lời nói, hay phản hồi khách hàng.

2.2.2. Quy trình hoạt động

Quy trình hoạt động của máy học thường bao gồm các bước sau:

1. Chuẩn bị dữ liệu:

- Thu thập dữ liệu: Thu thập dữ liệu từ các nguồn khác nhau, đảm bảo dữ liệu đa dạng và đầy đủ.
- Xử lý dữ liệu: Làm sạch và xử lý dữ liệu để loại bỏ nhiễu, xử lý các giá trị thiếu, và chuẩn hóa dữ liệu. Điều này đảm bảo rằng dữ liệu đầu vào có chất lượng tốt nhất.

2. Xây dựng mô hình:

- Lựa chọn thuật toán: Chọn các thuật toán máy học phù hợp với loại dữ liệu và bài toán cụ thể.

- Tiền xử lý đặc trưng: Chọn các đặc trưng quan trọng từ dữ liệu và tiền xử lý chúng để phù hợp với mô hình.

3. Huấn luyện mô hình:

- Chia dữ liệu: Chia dữ liệu thành tập huấn luyện và tập kiểm tra.
- Đào tạo mô hình: Sử dụng tập huấn luyện đã chuẩn bị để huấn luyện mô hình, điều chỉnh các tham số của hàm toán học để học các mối quan hệ giữa đầu vào và đầu ra.

4. Đánh giá mô hình:

- Kiểm tra mô hình: Sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình. Các chỉ số đánh giá có thể bao gồm độ chính xác, độ nhạy, độ đặc hiệu và các chỉ số khác tùy thuộc vào bài toán cụ thể.

5. Cải thiện mô hình:

- Tối ưu hóa: Điều chỉnh và tối ưu hóa mô hình để cải thiện hiệu suất. Việc này có thể bao gồm tinh chỉnh các siêu tham số hoặc thay đổi mô hình.
- Đánh giá lại: Lặp lại quá trình đánh giá để đảm bảo rằng các cải tiến đã đạt được kết quả mong muốn.

6. Triển khai mô hình:

- Đưa vào thực tế: Sau khi mô hình được tối ưu hóa, triển khai mô hình vào môi trường thực tế để đưa ra các dự đoán dựa trên dữ liệu mới.

19- Giám sát và cập nhật: Liên tục giám sát hiệu suất của mô hình trong môi trường thực tế và cập nhật mô hình khi cần thiết để đảm bảo nó vẫn hoạt động tốt với dữ liệu mới.



Hình 1. Quy trình hoạt động của học máy (Nguồn: internet)

2.2.3. Phân loại

Trong học máy, các phương pháp học được phân chia thành bốn nhóm chính, bao gồm: học có giám sát, học không giám sát, học bán giám sát và học tăng cường. Mỗi phương pháp có đặc điểm riêng biệt, phù hợp với từng loại dữ liệu và mục tiêu bài toán cụ thể.

2.2.3.1. Học có giám sát (Supervised Learning)

Đây là phương pháp trong đó mô hình được huấn luyện dựa trên tập dữ liệu đã được gán nhãn đầu ra. Mục tiêu là học một ánh xạ từ đầu vào đến đầu ra để dự đoán chính xác kết quả cho dữ liệu mới. Học có giám sát thường được sử dụng trong các bài toán phân loại và hồi quy, đồng thời có thể kết hợp với nhiều thuật toán khác nhau để tăng độ chính xác. Tuy nhiên, việc thu thập và gán nhãn cho một lượng lớn dữ liệu huấn luyện là một thách thức đáng kể về mặt chi phí và thời gian. Một số thuật toán phổ biến trong nhóm này gồm: mạng nơ-ron nhân tạo (ANN), Naive Bayes, hồi quy tuyến tính, hồi quy logistic, và máy vector hỗ trợ (SVM).

2.2.3.2. Học không giám sát (Unsupervised Learning)

Trong học không giám sát, mô hình được đào tạo trên tập dữ liệu không có nhãn, với mục tiêu phát hiện ra các mẫu tiềm ẩn hoặc cấu trúc nội tại trong dữ liệu. Phương pháp này đặc biệt hữu ích trong các tác vụ khám phá dữ liệu, phân cụm khách hàng, phát hiện bất thường, phân tích hành vi người dùng và nhận dạng mẫu. Các kỹ thuật nổi bật trong nhóm này bao gồm k-means clustering, phân tích thành phần chính (PCA) và phân tích giá trị kỳ dị (SVD).

2.2.3.3. Học bán giám sát (Semi-supervised Learning)

Học bán giám sát là sự kết hợp giữa học có giám sát và học không giám sát. Mô hình được huấn luyện trên một lượng nhỏ dữ liệu có gán nhãn cùng với một lượng lớn dữ liệu không gán nhãn. Cách tiếp cận này đặc biệt hiệu quả khi dữ liệu có nhãn khan hiếm hoặc tốn kém để thu thập. Học bán giám sát giúp cải thiện độ chính xác và khả năng khái quát của mô hình, đồng thời giảm số chiều đặc trưng và chi phí gán nhãn dữ liệu.

2.2.3.4. Học tăng cường (Reinforcement Learning)

Khác với ba phương pháp trên, học tăng cường dựa trên nguyên lý thử - sai, trong đó tác nhân (agent) tương tác với môi trường và học cách đưa ra quyết định thông qua cơ chế phần thưởng (reward) và phạt (penalty). Mục tiêu của mô hình là tối đa hóa tổng phần thưởng tích lũy theo thời gian. Phương pháp này đã đạt nhiều thành tựu nổi bật trong các lĩnh vực như chơi game (ví dụ: AlphaGo), robot học và điều khiển tự động. Tuy nhiên, việc triển khai trong thực tiễn doanh nghiệp vẫn còn hạn chế do độ phức tạp và tính khó đoán trong việc thiết kế phần thưởng tối ưu.

2.2.4. Các thuật toán học máy

2.2.4.1. Thuật toán Logistic Regression

Thuật toán Logistic Regression (Hồi quy Logistic) là một phương pháp học máy được sử dụng để dự đoán xác suất của một kết quả nhị phân (ví dụ: có/không, 0/1, tích cực/tiêu cực) dựa trên một hoặc nhiều biến số đầu vào (còn gọi là đặc trưng hoặc features). Dù tên gọi có chứa "regression" (hồi quy), nó thực chất được sử dụng cho bài toán phân loại, không phải dự đoán giá trị liên tục như hồi quy tuyến tính.

Logistic Regression sử dụng hàm sigmoid để ánh xạ giá trị đầu ra (tuyến tính) thành khoảng (0, 1), biểu thị xác suất. Công thức hàm sigmoid là:

$$y = f(s) = \frac{1}{1 + e^{-s}}$$

Trong đó:

$f(s)$ là xác suất xảy ra $y = 1$ hoặc $y = 0$

s là phương trình tuyến tính phụ thuộc vào biến đầu vào.

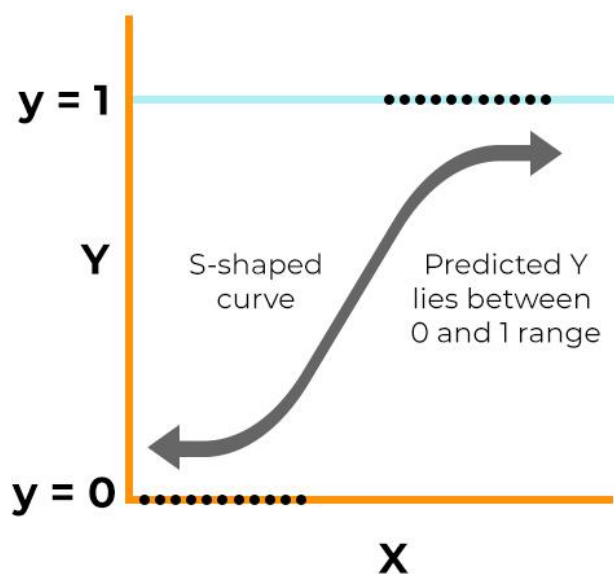
Phương trình mô hình đơn biến: $s = \alpha_0 + \alpha_1 x_1$, phương trình tuyến tính phụ thuộc vào duy nhất biến x_1 . Phương trình mô hình đa biến: $s = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_n x_n$, phương trình tuyến tính phụ thuộc vào các biến x .

Thuật toán tìm cách tối ưu hóa các tham số α sao cho mô hình dự đoán chính xác nhất, thường thông qua việc giảm thiểu hàm mất mát (log-loss hay cross-entropy loss).

Hàm mất mát: là hàm số xác định sự chênh lệch giữa đầu ra y dự đoán so với kết quả đầu ra y đã đúng (y dùng trong huấn luyện). Việc tối ưu hàm mất mát sẽ cho ra kết quả bài toán chính xác hơn.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^n -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Nó là một độ đo (metric) đo lường mức độ tương quan giữa phân phối xác suất dự báo và phân phối xác suất thực tế. Giá trị của hàm mất mát sẽ càng nhỏ nếu hai phân phối xác suất càng sát nhau, tức là giá trị dự báo giống với giá trị thực tế nhất.



Hình 2. Thuật toán Logistic Regression (Trích nguồn Internet)

2.2.4.2. Thuật toán Super Vector Machine

Support Vector Machine (SVM) là một phương pháp phân loại thống kê dựa trên việc tối đa hóa ranh giới giữa các điểm dữ liệu và siêu mặt phẳng phân tách. Thuật toán này định vị các biên tốt nhất để phân tách giữa dữ liệu tích cực và tiêu cực, được sử dụng rộng rãi vì hiệu suất vượt trội so với các phương pháp khác trong hầu hết các mô hình học máy.

Ý tưởng cốt lõi của SVM là xác định một mặt phẳng (hyper-plane) phù hợp để phân tách dữ liệu trong không gian nhiều chiều.

Công thức của mặt phẳng được biểu diễn như sau:

$$f(x) = w^T x + b$$

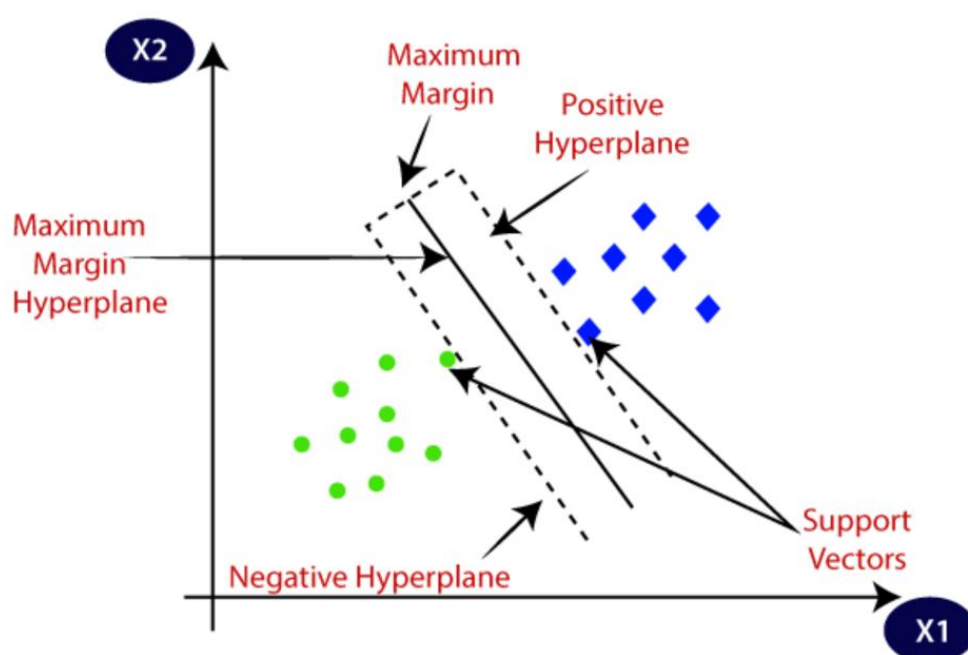
Trong đó:

w^T là ma trận chuyển vị

b là hệ số điều chỉnh

Đối với dữ liệu có tính tuyến tính (linear classifier), siêu phẳng là một đường thẳng: $y = ax + b$. Với dữ liệu phi tuyến tính, thuật toán SVM sẽ ánh xạ dữ liệu ban đầu sang không gian mới có số chiều là hữu hạn. Khi đó, đường ranh giới phân chia hai nhóm sẽ là một mặt phẳng trong không gian mới.

Mục tiêu chính là xác định khoảng cách (margin) lớn nhất giữa các lớp, nhằm giảm sai số khi phân loại dữ liệu mới. Điều này giúp tìm ra mặt phẳng có khoảng cách xa nhất so với các vector hỗ trợ. [6]



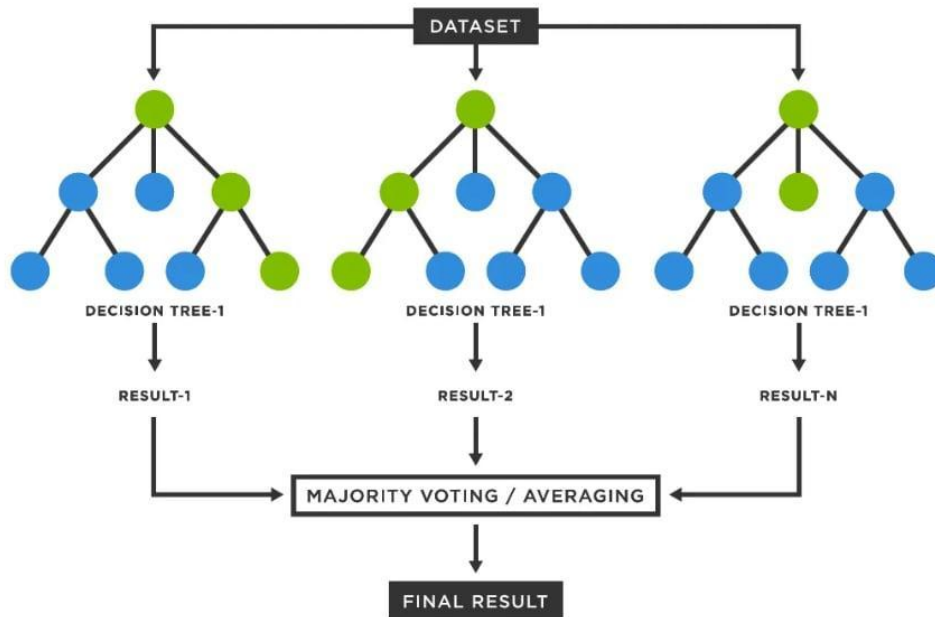
Hình 3. Thuật toán SVM (Trích nguồn Internet)

2.2.4.3. Thuật toán Random Forest

Random Forest là một thuật toán học máy có giám sát được sử dụng cho cả phân loại (Classification) và hồi quy (Regression). Nó là một phương pháp học tập tích hợp, sử dụng nhiều cây quyết định để huấn luyện và dự đoán. Thuật toán này kết hợp các khái

niệm về không gian con ngẫu nhiên bằng cách sử dụng nhiều cây quyết định làm yếu tố học và sử dụng phương pháp học tập tổng hợp bagging để kết hợp tất cả các kết quả dự đoán từ người học để có được kết quả dự đoán mạnh mẽ hơn

Thuật toán hoạt động dựa trên cách thức bỏ phiếu, có nghĩa là tất cả các cây quyết định trong mô hình sẽ thực hiện phân loại ra một lớp và lớp được bỏ phiếu nhiều nhất thì sẽ là đầu ra của mô hình.



Hình 4. Thuật toán Random Forest (Trích nguồn Internet)

2.2.4.4. Thuật toán XGBoost

XGBoost là một thuật toán tăng cường độ dốc (Gradient Boosting) hiệu suất cao, đặc biệt phù hợp với các bài toán phân loại văn bản lớn. Nó xây dựng mô hình bằng cách kết hợp nhiều cây quyết định yếu (weak learners) một cách tuần tự, với mỗi cây cố gắng sửa chữa các lỗi của các cây trước đó.

Công thức cập nhật trọng số trong Gradient Boosting:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

Trong đó:

$F_m(x)$ là mô hình sau khi thêm cây mới.

η là hệ số học (learning rate).

$h_m(x)$ là cây quyết định mới được thêm vào.

2.2.4.5. Thuật toán K-Nearest Neighbor

K-Nearest Neighbor (K-NN) là một thuật toán học máy không tham số được sử dụng cho phân loại (classification) và hồi quy (regression). Nguyên tắc chính của thuật toán là: Dự đoán nhãn của một điểm dữ liệu dựa trên nhãn của K điểm gần nhất trong không gian đặc trưng.

Cách hoạt động của K-NN:

Xác định giá trị K (số lượng hàng xóm gần nhất cần xem xét).

Tính khoảng cách giữa điểm cần dự đoán với tất cả các điểm trong tập dữ liệu huấn luyện. Thường dùng các công thức:

Khoảng cách Euclidean:

$$d(A, B) = \sqrt{\sum (A_i - B_i)^2}$$

Khoảng cách Manhattan:

$$d(A, B) = \sum |A_i - B_i|$$

Chọn K điểm gần nhất dựa trên khoảng cách đã tính.

Bỏ phiếu để dự đoán nhãn (đối với classification) hoặc tính trung bình (đối với regression).

2.2.4.6. Thuật toán Bernoulli Naive Bayes

Bernoulli Naive Bayes là một biến thể của thuật toán Naive Bayes, được xây dựng dựa trên định lý Bayes và giả định độc lập có điều kiện giữa các đặc trưng. Điểm đặc trưng của mô hình này là dữ liệu đầu vào được biểu diễn dưới dạng nhị phân – tức là mỗi đặc trưng chỉ có hai giá trị: có xuất hiện (1) hoặc không xuất hiện (0) trong văn bản.

(a) Nguyên lý hoạt động

Thuật toán sử dụng định lý Bayes để tính xác suất một văn bản xxx thuộc về một nhãn yyy, dựa trên tập hợp các đặc trưng xix_ixi, như sau:

$$P(y | x) = \frac{P(y) \prod_{i=1}^n P(x_i | y)}{P(x)}$$

Trong đó:

$P(y)$ là xác suất tiên nghiệm của lớp yyy

$P(x_i | y)$ là xác suất đặc trưng xix_ixi xuất hiện trong lớp yyy,

$P(x)$ là xác suất của dữ liệu đầu vào (thường được bỏ qua trong phân loại vì không phụ thuộc vào lớp).

Bernoulli Naive Bayes đặc biệt sử dụng đặc trưng nhị phân, nghĩa là mỗi từ trong từ vựng chỉ được xét là có xuất hiện hay không trong văn bản, không xét đến tần suất.

(b) Ứng dụng trong phân loại văn bản

Mô hình Bernoulli Naive Bayes được áp dụng rộng rãi trong các tác vụ:

Phân loại email rác (spam detection)

Phân tích cảm xúc (sentiment analysis)

Phân loại tin tức theo chủ đề

Trong các bài toán này, văn bản thường được biểu diễn bằng tập hợp các từ khóa (bag-of-words) với giá trị nhị phân, phù hợp với giả định của mô hình.

(c) Ưu điểm và hạn chế

Ưu điểm: Đơn giản, nhanh và dễ huấn luyện. Hiệu quả với dữ liệu có độ thưa cao (sparse data), như văn bản. Yêu cầu ít tài nguyên tính toán, phù hợp với mô hình có số lượng đặc trưng lớn.

Hạn chế: Giả định độc lập giữa các đặc trưng không thực tế trong ngôn ngữ tự nhiên. Không tận dụng thông tin về tần suất từ (so với mô hình Multinomial Naive Bayes). Có thể bị suy giảm hiệu suất nếu văn bản dài chứa nhiều từ không liên quan.

2.3. Học sâu

2.3.1. Tổng quan về học sâu

Học sâu (Deep Learning) là một nhánh của học máy (Machine Learning), trong đó các mô hình có cấu trúc nhiều tầng (deep neural networks) được sử dụng để trích xuất và học các đặc trưng ở nhiều cấp độ từ dữ liệu đầu vào. Khác với các phương pháp truyền thống yêu cầu kỹ thuật thủ công để chọn lựa đặc trưng (feature engineering), học sâu cho phép hệ thống tự động học đặc trưng một cách hiệu quả từ dữ liệu thô, nhờ vào khả năng biểu diễn phi tuyến mạnh mẽ của mạng nhiều lớp.

Học sâu đã đạt được những thành tựu vượt bậc trong nhiều lĩnh vực như thị giác máy tính, nhận diện giọng nói, xử lý ngôn ngữ tự nhiên (NLP) và xe tự hành. Sự phát triển mạnh mẽ của học sâu được thúc đẩy bởi ba yếu tố chính: (1) sự gia tăng đáng kể của dữ liệu, (2) sự cải thiện về phần cứng (GPU, TPU), và (3) các thuật toán tối ưu hóa hiệu quả.

2.3.2. Kiến trúc mạng nơ-ron nhân tạo

Mạng nơ-ron nhân tạo (Artificial Neural Network – ANN) là nền tảng của học sâu, được thiết kế mô phỏng cách thức hoạt động của các nơ-ron sinh học trong não người. Một mạng ANN bao gồm ba loại lớp chính:

Lớp đầu vào (Input Layer): nhận dữ liệu đầu vào ban đầu.

Lớp ẩn (Hidden Layers): xử lý và trích xuất đặc trưng. Số lượng lớp ẩn và số nơ-ron trong mỗi lớp xác định "độ sâu" của mạng.

Lớp đầu ra (Output Layer): đưa ra dự đoán cuối cùng, thường tương ứng với một hoặc nhiều nhãn.

Mỗi nơ-ron trong một lớp được kết nối với tất cả các nơ-ron ở lớp trước và sau, thông qua các trọng số (weights). Trong quá trình huấn luyện, các trọng số này được điều chỉnh để tối thiểu hóa sai số dự đoán.

2.3.3. Hàm kích hoạt

Hàm kích hoạt (Activation Function) là thành phần quyết định tính phi tuyến của mạng nơ-ron, cho phép mô hình học được các quan hệ phức tạp trong dữ liệu. Một số hàm kích hoạt phổ biến bao gồm:

Sigmoid: Dễ hiểu, nhưng gây ra hiện tượng mất dần gradient (vanishing gradient).

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Tanh: Giảm thiểu bù bias, nhưng vẫn gặp vấn đề mất gradient.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

ReLU (Rectified Linear Unit): Đơn giản, hiệu quả và thường được dùng trong mạng sâu hiện đại.

2.3.4. Kiến trúc học sâu trong NLP

Trong xử lý ngôn ngữ tự nhiên, dữ liệu đầu vào thường ở dạng văn bản. Các kiến trúc học sâu giúp biểu diễn và xử lý ngôn ngữ theo cách mà các mô hình truyền thống không làm được, nhờ khả năng học ngữ cảnh và trật tự chuỗi từ.

2.3.4.1. Long Short-Term Memory - LSTM

LSTM là một biến thể đặc biệt của mạng nơ-ron hồi tiếp (RNN – Recurrent Neural Network), được thiết kế nhằm khắc phục vấn đề gradient biến mất (vanishing gradient problem) trong quá trình huấn luyện RNN truyền thống khi xử lý các chuỗi dài.

Trong các ứng dụng xử lý ngôn ngữ tự nhiên (NLP), các chuỗi dữ liệu (câu, đoạn văn) thường có độ dài khác nhau, và thông tin ở thời điểm trước đó có thể quan trọng để đưa ra dự đoán ở thời điểm hiện tại. Tuy nhiên, RNN thông thường gặp khó khăn trong việc ghi nhớ thông tin lâu dài. LSTM được thiết kế để giải quyết vấn đề này thông qua một cơ chế bộ nhớ có khả năng lưu giữ thông tin quan trọng trong thời gian dài.

Cấu trúc của một tế bào LSTM: Mỗi tế bào (cell) trong LSTM có ba thành phần chính được gọi là các cổng (gates):

* Cổng quên (Forget Gate):

Công thức:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Mục tiêu: xác định phần nào của thông tin từ trạng thái ẩn trước đó h_{t-1} và đầu vào hiện tại x_t cần được loại bỏ khỏi bộ nhớ.

Nếu f_0 gần 0 \rightarrow thông tin bị lãng quên; nếu gần 1 \rightarrow thông tin được giữ lại.

* Cổng đầu vào (Input Gate):

Công thức:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Mục tiêu: quyết định thông tin mới nào sẽ được thêm vào bộ nhớ.

Gồm hai bước:

i_t : chọn lọc thông tin nào sẽ được cập nhật.

C_t : tạo vector ứng viên giá trị mới để thêm vào trạng thái bộ nhớ.

* Cổng đầu ra (Output Gate):

Công thức:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \cdot \tanh(C_t)$$

Mục tiêu: xác định thông tin nào từ trạng thái bộ nhớ hiện tại sẽ được sử dụng để tính đầu ra tại thời điểm hiện tại.

* Cập nhật trạng thái bộ nhớ:

Bộ nhớ được cập nhật theo:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

Giúp lưu giữ cả thông tin cũ (qua f_t) và cập nhật thông tin mới (qua i_t).

2.3.4.2. CNN

Dù CNN nổi tiếng trong xử lý ảnh, nhưng cũng được sử dụng hiệu quả trong NLP để trích xuất đặc trưng cục bộ từ chuỗi văn bản. CNN có khả năng học các mẫu

(patterns) như cụm từ đặc trưng hoặc cụm cảm xúc, bất kể vị trí xuất hiện trong văn bản. Trong NLP, các bộ lọc (filter) hoạt động như lớp trượt qua câu, học ra các đặc trưng quan trọng cho phân loại.

2.4. Các phương pháp trích xuất đặc trưng

Trích xuất đặc trưng (Feature Extraction) là quá trình chuyển đổi dữ liệu thô (như văn bản, hình ảnh, âm thanh) thành dạng có thể xử lý bởi mô hình học máy. Chúng giúp giảm chiều dữ liệu và giữ lại thông tin quan trọng nhất để cải thiện hiệu suất của mô hình [4]. Có 2 dạng trích xuất đặc trưng cho văn bản:

Vector hóa (Vectorization): Biểu diễn văn bản bằng ma trận tần suất hoặc trọng số.

Ví dụ: Bag-of-Words, TF-IDF, N-grams, ...

Nhúng từ (Word Embedding): Học vector biểu diễn ngữ nghĩa của từ dựa trên ngữ cảnh.

Ví dụ: Word2Vec, Doc2Vec, GloVe, ...

2.4.1. Bag-of-words

Bag-of-Words, viết tắt là BoW, có nghĩa là túi từ. Mô hình BoW chỉ tập hợp tất cả các từ dạng một từ duy nhất, không chứa các cụm từ gồm nhiều từ ghép lại. Mô hình Bag of N-Grams sẽ giải quyết vấn đề này. Bag of N-grams sẽ thành lập một tập hợp các cụm từ gồm n-từ ghép lại với nhau tùy thuộc vào nhu cầu [3]. Các bước xây dựng thuật toán Bag-of-Words:

Bước 1 - Xây dựng tập từ vựng: Tạo một danh sách tất cả các từ duy nhất có trong tập dữ liệu văn bản.

Bước 2 - Chuyển đổi thành vector: Chuyển đổi mỗi văn bản thành một vector, trong đó mỗi phần tử của vector đại diện cho tần suất xuất hiện của một từ cụ thể trong từ điển. Độ dài của vector bằng với số lượng từ trong từ điển.

Ví dụ: Giả sử ta có 3 bình luận sau:

1. “Tôi thích đôi giày này”
2. “Đôi giày xinh quá”

3. “Giày không giống hình”

Bước 1: Xây dựng tập từ vựng:

Bình luận 1 có tập từ: “tôi thích”, “thích đôi”, “đôi giày”, “giày này”

Bình luận 2 có tập từ: “đôi giày”, “giày xinh”, “xinh quá”

Bình luận 3 có tập từ: “giày không”, “không giống”, “giống hình”

Bước 2: Chuyển đổi văn bản thành vector

Bình luận	Cụm từ								
	đôi giày	giày không	giày này	giày xinh	giống hình	không giống	thích đôi	tôi thích	xinh quá
1	1	0	1	0	0	0	1	1	0
2	1	0	0	1	0	0	0	0	1
3	0	1	0	0	1	1	0	0	0

Hình 5. Ma trận từ theo mô hình Bag-of-Words

2.4.2. TF-IDF

TF-IDF (Term Frequency-Invert Document Frequency) là một phương pháp thống kê được dùng để đo lường tầm quan trọng của một từ trong một tập tài liệu, giúp đánh giá mức độ liên quan của một từ dựa trên tần suất xuất hiện của nó trong một văn bản, đồng thời giảm thiểu tầm quan trọng của các từ phổ biến xuất hiện trong nhiều văn bản. TF-IDF là một phương thức vector hóa văn bản, văn bản sẽ được biểu diễn thành dạng vector làm đầu vào cho mô hình học máy. [4]

Công thức tính toán TF-IDF bao gồm hai thành phần chính:

TF (Term Frequency - Tần số thuật ngữ): Đo tần suất một từ xuất hiện trong một văn bản.

$$TF(t, d) = \frac{f(t, d)}{T}$$

Trong đó:

t là một từ trong văn bản.

f(t,d) là tần suất của t xuất hiện trong đoạn văn bản d.

T là tổng số từ của đoạn văn bản.

IDF (Inverse Document Frequency - Tần số tài liệu nghịch đảo): Đo mức độ quan trọng của một từ trong toàn bộ tập tài liệu.

$$IDF(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Trong đó:

N: tổng số văn bản trong tập D.

Mẫu số là số văn bản có chứa từ t.

Nếu từ đó không xuất hiện ở bất cứ một văn bản nào trong tập dữ liệu thì mẫu số sẽ bằng 0, phép chia cho 0 là phép chia không hợp lệ. Vì thế với trường hợp này, 1 được cộng thêm vào mẫu số. TF-IDF được tính như sau:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

Những từ có giá trị TF-IDF cao là những từ xuất hiện trong văn bản này và ít xuất hiện ở văn bản khác. Do đó, việc này giúp loại bỏ những từ xuất hiện nhiều nhưng ít có ý nghĩa và giữ lại những từ có giá trị cao.

2.4.3. N-grams

N-grams là một kỹ thuật cơ bản nhưng quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), được sử dụng để biểu diễn văn bản dưới dạng chuỗi các từ hoặc ký tự liên tiếp. Nói cách khác, N-grams là các tập hợp con có thứ tự gồm N phần tử (thường là từ hoặc ký tự) được trích xuất từ một đoạn văn bản, giúp mô hình máy học nắm bắt được ngữ cảnh cục bộ trong dữ liệu ngôn ngữ.

2.4.3.1. Khái niệm và phân loại

Một N-gram là một chuỗi liên tiếp gồm N đơn vị ngôn ngữ. Tùy thuộc vào giá trị của N, ta có:

Unigram (1-gram): các đơn vị đơn lẻ (thường là từ hoặc ký tự).

Bigram (2-gram): cặp hai từ/ký tự liên tiếp.

Trigram (3-gram): bộ ba từ/ký tự liên tiếp.

...

N-gram: chuỗi N từ/ký tự liên tục trong văn bản.

Ví dụ, với câu: *"Tôi thích học máy"*, các N-grams thu được là:

Unigrams: ["Tôi", "thích", "học", "máy"]

Bigrams: ["Tôi thích", "thích học", "học máy"]

Trigrams: ["Tôi thích học", "thích học máy"]

2.4.3.2. Ứng dụng và vai trò

N-grams đóng vai trò quan trọng trong nhiều tác vụ xử lý ngôn ngữ như: Mô hình ngôn ngữ (Language Modeling): xác định xác suất xuất hiện của một chuỗi từ dựa trên các từ trước đó. Phân loại văn bản (Text Classification): chuyển văn bản thành đặc trưng đầu vào cho các mô hình học máy. Phân tích cảm xúc (Sentiment Analysis): nắm bắt các cụm từ biểu thị thái độ tích cực hoặc tiêu cực. Phát hiện đạo văn, kiểm tra chính tả, dịch máy, v.v.

2.4.3.3. Ưu điểm và hạn chế

Ưu điểm: Dễ triển khai, tính toán nhanh. Nắm bắt được mối liên hệ cục bộ giữa các từ. Hữu ích trong các mô hình tuyến tính hoặc khi lượng dữ liệu không quá lớn.

Hạn chế: Số lượng đặc trưng tăng nhanh khi N tăng → "hiện tượng nổ chiều" (curse of dimensionality). Thiếu khả năng mô hình hóa ngữ cảnh dài hạn. Không hiểu được ý nghĩa ngôn ngữ học sâu (như cú pháp, ngữ nghĩa).

Để khắc phục các hạn chế trên, các mô hình hiện đại như Word2Vec, BERT, Transformer ra đời, cho phép biểu diễn từ dưới dạng vector ngữ nghĩa, nắm bắt được cả mối quan hệ ngữ cảnh gần và xa. Tuy nhiên, N-grams vẫn là công cụ tiền xử lý quan trọng trong các hệ thống học máy truyền thống và là nền tảng giúp hiểu rõ hơn các đặc trưng ngôn ngữ.

2.5. Phương pháp đánh giá mô hình

Trong bài toán phân loại, việc lựa chọn các chỉ số đánh giá phù hợp đóng vai trò quan trọng nhằm phản ánh hiệu quả của mô hình một cách khách quan. Hai thước đo phổ biến thường được sử dụng là Accuracy và ROC-AUC.

2.5.1. Accuracy

Accuracy là một trong những chỉ số cơ bản và phổ biến nhất dùng để đánh giá hiệu suất của mô hình phân loại. Chỉ số này được tính bằng tỷ lệ giữa số lượng dự đoán chính xác (bao gồm cả các trường hợp dương tính đúng – *True Positive* và âm tính đúng – *True Negative*) so với tổng số mẫu trong tập dữ liệu. Cụ thể, công thức tính Accuracy được biểu diễn như sau:

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

Trong đó:

TP (True Positive): Số lượng mẫu thực sự dương tính và được mô hình dự đoán đúng.

TN (True Negative): Số lượng mẫu thực sự âm tính và được mô hình dự đoán đúng.

FP (False Positive): Số lượng mẫu thực sự âm tính nhưng bị dự đoán nhầm là dương tính.

FN (False Negative): Số lượng mẫu thực sự dương tính nhưng bị dự đoán nhầm là âm tính.

Accuracy cung cấp một cái nhìn tổng quát về mức độ đúng đắn của mô hình trong việc phân loại các mẫu. Tuy nhiên, trong các bài toán có dữ liệu không cân bằng – ví dụ như khi một lớp chiếm ưu thế tuyệt đối so với lớp còn lại – thì độ chính xác có thể trở nên không phản ánh đúng hiệu quả thực sự của mô hình. Chẳng hạn, nếu 95% dữ liệu thuộc về lớp âm tính, một mô hình luôn dự đoán âm tính sẽ đạt Accuracy 95%, dù không có khả năng nhận diện lớp dương tính. Do đó, trong những trường hợp như vậy, cần bổ sung các chỉ số khác như Precision, Recall, F1-score hoặc ROC-AUC để có cái nhìn toàn diện hơn về hiệu suất phân loại.

2.5.2. ROC-AUC

ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) là một chỉ số đánh giá quan trọng trong các bài toán phân loại nhị phân, đặc biệt hữu ích trong bối cảnh dữ liệu không cân bằng giữa các lớp. ROC (Receiver Operating Characteristic) là một đường cong thể hiện mối quan hệ giữa True Positive Rate (TPR) và False Positive Rate (FPR) ở các ngưỡng phân loại khác nhau.

True Positive Rate (TPR), còn gọi là *Recall* hoặc *Sensitivity*, được tính theo công thức:

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

False Positive Rate (FPR) được tính như sau:

$$FPR = \frac{FP}{FP + TN}$$

Khi thay đổi ngưỡng quyết định (threshold) mà mô hình sử dụng để phân loại đầu ra thành dương tính hay âm tính, các giá trị TPR và FPR sẽ thay đổi tương ứng. Việc vẽ TPR theo FPR ở nhiều mức ngưỡng khác nhau sẽ tạo ra đường cong ROC.

AUC (Area Under the Curve) là phần diện tích dưới đường cong ROC và được dùng như một thước đo tổng quát cho khả năng phân biệt giữa hai lớp của mô hình. Giá trị AUC nằm trong khoảng từ 0 đến 1, với các ý nghĩa như sau:

AUC = 1.0: Mô hình phân biệt hoàn hảo giữa hai lớp.

AUC = 0.5: Mô hình không có khả năng phân biệt, tương đương với việc đoán ngẫu nhiên.

AUC < 0.5: Mô hình hoạt động tệ hơn cả đoán ngẫu nhiên, có thể đã bị huấn luyện sai hướng.

Ưu điểm chính của ROC-AUC là nó không phụ thuộc vào tỷ lệ phân bố lớp trong tập dữ liệu, nên thường cho kết quả đánh giá đáng tin cậy hơn Accuracy trong các bài toán có hiện tượng mất cân bằng lớp (*class imbalance*). Ví dụ, trong một tập dữ liệu mà 90% mẫu thuộc về lớp âm tính và chỉ 10% mẫu thuộc về lớp dương tính, mô hình có thể đạt Accuracy cao nhưng lại thất bại trong việc phát hiện lớp thiểu số. ROC-AUC, trong trường hợp này, sẽ giúp đánh giá liệu mô hình có thật sự học được đặc trưng phân biệt hai lớp hay không.

2.6. Django Framework:

2.6.1. Tổng quan về Django

Django là một framework mã nguồn mở được viết bằng ngôn ngữ Python, dùng để phát triển các ứng dụng web theo mô hình MTV (Model - Template - View). Django được thiết kế nhằm giúp lập trình viên phát triển nhanh chóng, an toàn và có khả năng mở rộng tốt, đồng thời tuân thủ nguyên tắc DRY (Don't Repeat Yourself) – hạn chế việc lặp lại mã không cần thiết.

Django được phát triển lần đầu vào năm 2005 bởi một nhóm các nhà phát triển web tại Lawrence Journal-World, với mục tiêu cung cấp một giải pháp toàn diện cho việc xây dựng các ứng dụng web hiện đại.

2.6.2. Kiến trúc MTV (Model - Template - View)

Mặc dù giống mô hình MVC (Model-View-Controller), Django tổ chức theo mô hình MTV:

Model: Đại diện cho tầng dữ liệu – định nghĩa cấu trúc cơ sở dữ liệu, các trường (fields) và hành vi (methods) thông qua các lớp Python. Django sử dụng hệ ORM (Object Relational Mapping) để ánh xạ giữa database và code.

Template: Tầng giao diện – nơi hiển thị dữ liệu cho người dùng cuối, sử dụng hệ thống template của Django để kết hợp HTML với các biến và logic hiển thị.

View: Điều phối logic xử lý – chịu trách nhiệm nhận yêu cầu (request) từ người dùng, xử lý dữ liệu từ model và trả về kết quả qua template.

2.6.3. Các thành phần chính trong Django

Thành phần	Vai trò
URL Dispatcher	Ánh xạ đường dẫn URL tới các hàm xử lý tương ứng (views).
ORM (Object-Relational Mapping)	Cho phép tương tác với cơ sở dữ liệu thông qua các đối tượng Python, thay vì SQL thuần.

Thành phần	Vai trò
Forms	Hỗ trợ tạo, xử lý và xác thực biểu mẫu (form) trong HTML.
Admin Interface	Giao diện quản trị tự động, dễ tùy biến, dùng để quản lý dữ liệu trong ứng dụng.
Middleware	Tầng xử lý trung gian giữa request và response, dùng cho các chức năng như xác thực, cache, logging...
Authentication System	Tích hợp sẵn hệ thống quản lý người dùng, đăng nhập, phân quyền.

Bảng 1. Các thành phần chính trong Django.

2.6.4. Ưu điểm của Django

Phát triển nhanh: Cung cấp nhiều thành phần tích hợp sẵn như admin panel, xác thực người dùng, routing...

Bảo mật cao: Django xử lý nhiều lỗ hổng phổ biến như CSRF, SQL Injection, XSS...

Khả năng mở rộng tốt: Dễ tích hợp với REST API, Redis, Docker, và các công nghệ hiện đại.

Cộng đồng lớn: Tài liệu đầy đủ, thư viện phong phú (Django REST Framework, Celery, Channels...).

Chương 3. HỆ THỐNG DỰ ĐOÁN CẢM XÚC

3.1. Giới thiệu hệ thống

Hệ thống dự đoán cảm xúc được phát triển trong đề tài này có nhiệm vụ tự động phân tích các đánh giá văn bản và xác định cảm xúc chủ đạo là tích cực hay tiêu cực. Hệ thống kết hợp giữa các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và mô hình học máy/học sâu, cho phép máy tính hiểu và phân loại ngữ nghĩa của ngôn ngữ tự nhiên một cách chính xác.

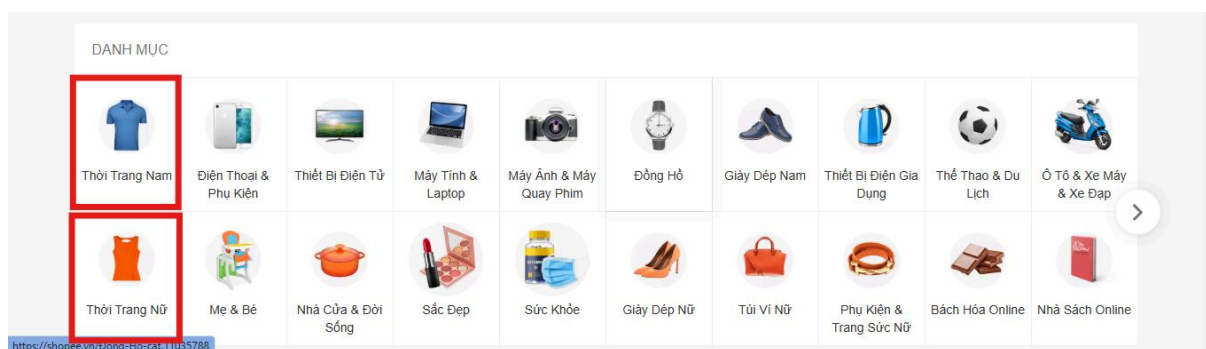
Về mặt tổng thể, hệ thống gồm các bước chính: tiền xử lý văn bản, vector hóa dữ liệu đầu vào, dự đoán cảm xúc bằng mô hình học máy đã huấn luyện, và hiển thị kết quả cho người dùng thông qua giao diện web. Giao diện được xây dựng bằng framework Django, cho phép người dùng dễ dàng nhập nội dung đánh giá và nhận được kết quả phân tích nhanh chóng, chính xác.

Việc tự động hóa quá trình đánh giá cảm xúc không chỉ giúp tiết kiệm thời gian và chi phí so với việc phân tích thủ công, mà còn hỗ trợ doanh nghiệp đưa ra các quyết định cải tiến sản phẩm và dịch vụ dựa trên dữ liệu thực tiễn từ khách hàng.

3.2. Kiến trúc hệ thống

3.2.1. Thu thập dữ liệu

Bộ dữ liệu được lấy trực tiếp từ website TMĐT Shopee bằng cách xây dựng một đoạn code Python sử dụng thư viện BeautifulSoup và API của Shopee. Dữ liệu được crawl về là các bình luận và đánh giá của khách hàng về các sản phẩm phổ biến thuộc nhóm ngành thời trang, gồm 7 danh mục: thời trang nam, thời trang nữ.



Hình 6. Danh mục dữ liệu được thu thập

Bộ dữ liệu bao gồm 221276 dòng và 2 cột thông tin. Các dòng này chứa nhận xét từ khách hàng về sản phẩm thuộc danh mục thời trang trên Shopee. Được mô tả chi tiết như sau:

Tên cột	Ý nghĩa
rating	Mức độ phản hồi về chất lượng sản phẩm từ khách hàng, với thang đo từ 1 sao đến 5 sao.
comment	Nội dung phản hồi dưới dạng văn bản do khách hàng cung cấp. Nhận xét này có thể liên quan đến chất lượng sản phẩm, dịch vụ khách hàng, thời gian giao hàng, ...

Bảng 2. Mô tả các thuộc tính của bộ dữ liệu.

3.2.2. Đào tạo mô hình

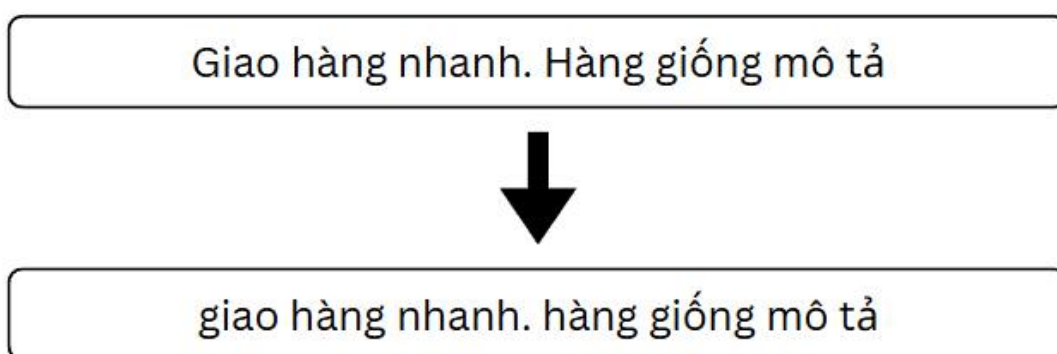
3.2.2.1. Tiền xử lý dữ liệu

Dữ liệu thu thập ban đầu thường ở trạng thái thô, chưa được xử lý, và có thể tồn tại các vấn đề như giá trị thiếu, dữ liệu trùng lặp hoặc lỗi chính tả. Bên cạnh đó, nội dung dữ liệu đôi khi chứa các thành phần không mong muốn như liên kết quảng cáo hoặc ký tự đặc biệt. Những yếu tố này có thể làm sai lệch kết quả phân tích và ảnh hưởng đến hiệu suất của mô hình. Vì vậy, quá trình tiền xử lý và làm sạch dữ liệu là bước quan trọng nhằm đảm bảo chất lượng đầu vào cho mô hình học máy.

a) Xóa dữ liệu rỗng và dữ liệu trùng

Tập dữ liệu có 99675 hàng rỗng và 16158 hàng trùng lặp. Đây có thể là sai sót khi tiến hành cào dữ liệu dẫn đến tình trạng trên. Đương nhiên những dữ liệu trên không mang giá trị phân tích nên sẽ được xóa khỏi tập dữ liệu. Sau cùng, bộ dữ liệu còn 105442 hàng.

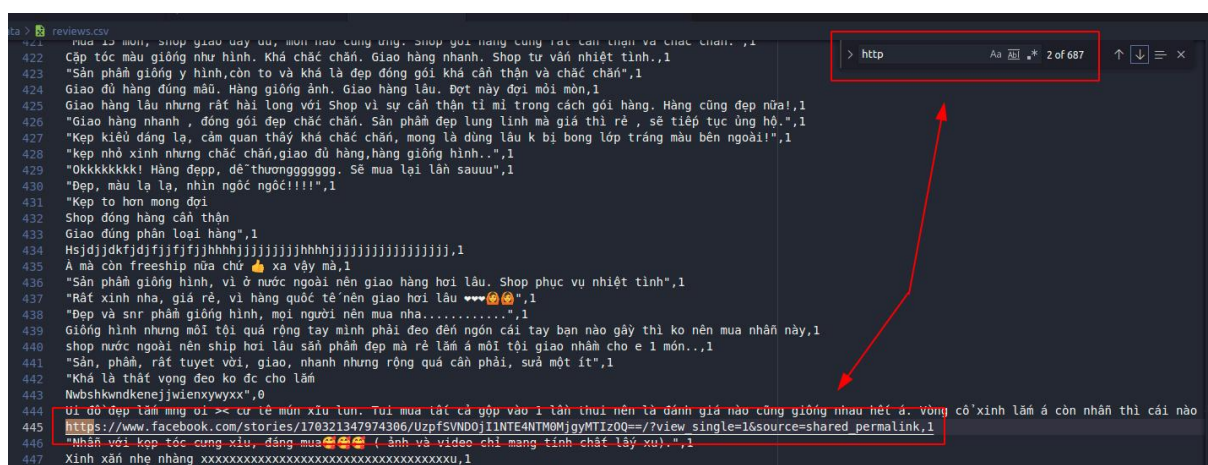
b) Chuyển các đánh giá thành chữ thường và cùng một chuẩn Unicode (NFD):



Hình 7. Minh họa chuẩn mã hóa dữ liệu

Trong quá trình thu thập dữ liệu văn bản, đặc biệt là bình luận của người dùng, có thể xuất hiện sự không đồng nhất về bảng mã ký tự, chẳng hạn như TCVN3, VNI, VPS, UTF-8, UTF-16,... Sự đa dạng về mã hóa này khiến cho cùng một từ hoặc cụm từ có thể được biểu diễn dưới nhiều dạng khác nhau trong dữ liệu. Nếu không thực hiện chuẩn hóa bảng mã về một định dạng thống nhất (thường là UTF-8), hệ thống phân tích sẽ hiểu đây là những chuỗi hoàn toàn khác biệt, dẫn đến sai lệch trong quá trình xử lý và giảm độ chính xác của mô hình. Do đó, chuẩn hóa mã hóa văn bản là một bước tiền xử lý quan trọng nhằm đảm bảo tính nhất quán và độ tin cậy của dữ liệu đầu vào.

c) Xóa đánh giá chứa link url



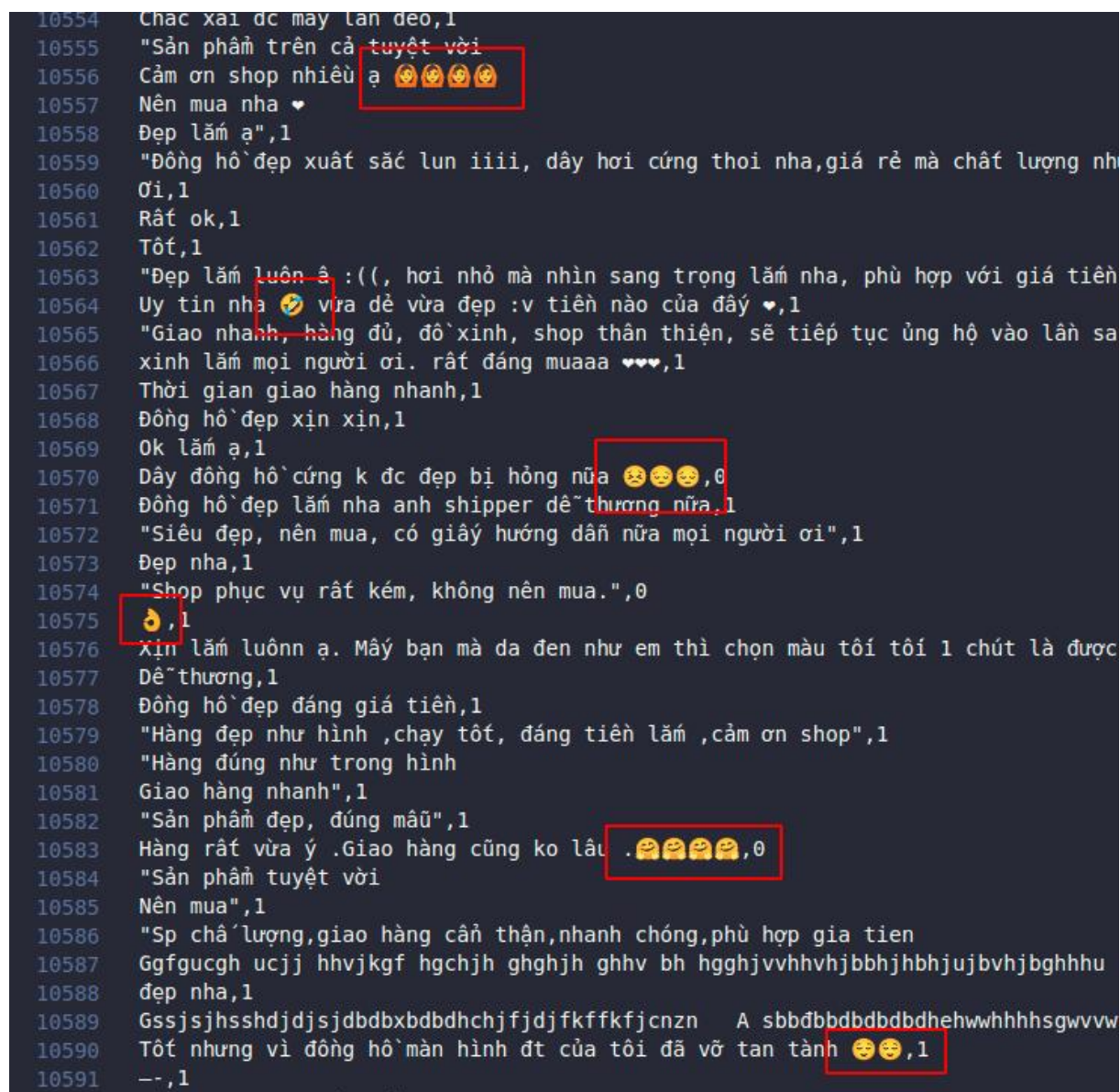
Hình 8. Đánh giá chứa link url.

Trên nền tảng thương mại điện tử Shopee, một số bình luận sản phẩm được người dùng đăng tải chủ yếu nhằm mục đích nhận xu thưởng, thay vì thể hiện ý kiến

đánh giá thực chất về sản phẩm. Nhiều trong số đó chứa các đường dẫn quảng cáo hoặc nội dung sao chép từ nguồn khác chỉ để đạt đủ số ký tự cần thiết (Hình 3-5). Những bình luận như vậy không phản ánh cảm xúc hoặc trải nghiệm thực tế của người dùng, do đó được xem là nhiễu và cần được loại bỏ trong quá trình tiền xử lý dữ liệu để đảm bảo chất lượng đầu vào cho mô hình phân tích cảm xúc.

d) Xử lý ký tự đặc biệt

Khi quan sát các bình luận, có thể nhận thấy nhiều trường hợp chứa biểu tượng cảm xúc (emoji), như minh họa trong hình dưới đây:

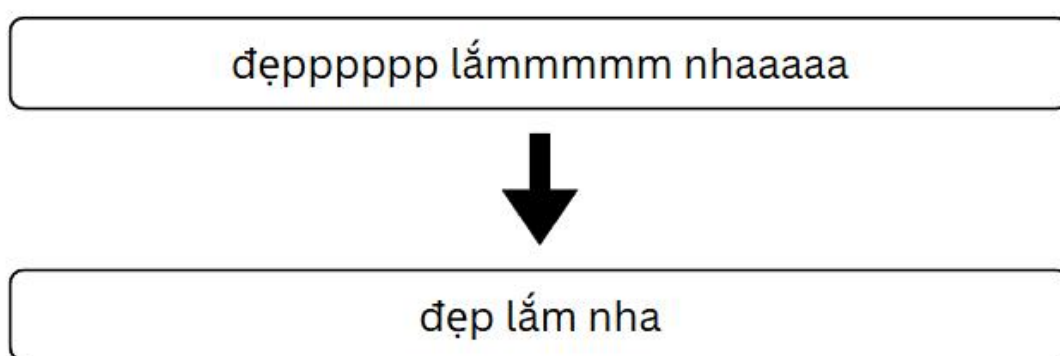


Hình 9. Các emojis trong đánh giá.

Emoji là một yếu tố quan trọng góp phần tăng cường khả năng nhận diện cảm xúc của mô hình phân tích. Tuy nhiên, nếu trong quá trình tiền xử lý, ta thực hiện loại bỏ các ký tự đặc biệt trước khi xử lý emoji, thì có nguy cơ các emoji này cũng sẽ bị loại bỏ do chúng vốn được mã hóa từ các ký tự đặc biệt. Điều này dẫn đến việc mất đi một nguồn thông tin cảm xúc có giá trị. Như hình minh họa, rõ ràng emoji có thể giúp làm rõ ngữ nghĩa và sắc thái cảm xúc (tích cực hoặc tiêu cực) của một bình luận, từ đó hỗ trợ mô hình đưa ra dự đoán chính xác hơn.

Do đó, hướng xử lý các ký tự đặc biệt lần lượt thực hiện như sau: tách các emojis sang một cột mới, sau đó tiến hành xóa các ký tự đặc biệt khác.

e) Xử lý từ lặp ký tự cuối



Hình 10. Xử lý các từ bị lặp ký tự cuối.

Cần chuẩn hóa các từ bị kéo dài ký tự lặp lại như: "chờiiiiii ơiiiiii", "xinhhhhhhh quá", "đẹp xiuuuuuuuuuuuuu" về dạng chuẩn như "chời ơi", "xinh quá", "đẹp xiu". Việc này giúp loại bỏ nhiễu trong văn bản và đảm bảo tính nhất quán trong tập dữ liệu.

Tuy nhiên, một vấn đề cần lưu ý là khi áp dụng kỹ thuật này một cách không kiểm soát, các từ tiếng Anh hợp lệ như "feedback" có thể bị biến dạng thành "fedback", gây mất ý nghĩa ban đầu. Do đó, quá trình chuẩn hóa ký tự lặp sẽ được thực hiện theo trình tự: kiểm tra có phải tiếng anh không? -> tiến hành loại bỏ các chữ bị lặp.

f) Xóa các đánh giá vô nghĩa

Trên các nền tảng thương mại điện tử nói chung, và Shopee nói riêng, thường xuất hiện nhiều bình luận không mang ý nghĩa rõ ràng, như minh họa trong Hình 3-7. Những bình luận này chứa các từ không có trong từ điển tiếng Việt, được xem là

từ vô nghĩa và không đóng góp thông tin hữu ích cho quá trình phân tích. Do đó, cần thực hiện bước xử lý nhằm loại bỏ các bình luận dạng này bằng cách xây dựng một từ điển chứa toàn bộ các từ đơn hợp lệ trong tiếng Việt, từ đó đối chiếu và lọc ra những bình luận không hợp lệ để loại bỏ khỏi tập dữ liệu.

```

8519 9152588873,Đồng Hồ,truongxuan075,5,Ok hàng đúng theo mô tả giao hàng nhanh chóng rất thích và hài lòng
8520 9152588873,Đồng Hồ,l*****8,5,sản phẩm tốt lần sau sẽ ủng hộ shop tiếp. Cho shop 5 sao. Good
8521 9152588873,Đồng Hồ,cthanh_1311,5,hàng giao đẹp lắm shop ơi em rất hài lòng với sản phẩm bên mình
8522 9152588873,Đồng Hồ,o*****3,5,Vải khá là ok nhìn đẹp đấy nhé nhé tuyệt vời lắm em qidndkabfjf
8523 9152588873,Đồng Hồ,nh1408,5,"Chất liệu:cao su
8524 Đúng với mô tả:đúng
8525 Màu sắc:đẹp
8526
8527 Quá là oke
8528 Hahsjsjsjejejejejrjrjtiekwjshsgsgshhshshshsjsj"
8529 9152588873,Đồng Hồ,5fmmu7s_35,5,"Chất liệu:Dây da, viền thép
8530 Màu sắc:Đen trắng
8531
8532 Đồng hồ đẹp, đeo đầm tay"

```

Hình 11. Đánh giá vô nghĩa.

g) Xử lý từ viết tắt

Giới trẻ thường dễ dàng tiếp cận các xu hướng công nghệ hơn. Shopee cũng thế, phần lớn khách hàng họ là những người trẻ. Họ mua sắm trực tuyến nhiều và cũng có xu hướng để lại đánh giá sản phẩm. Do đó, theo quan sát, trong tập dữ liệu có rất nhiều từ viết tắt. Việc thay thế các từ viết tắt này thực sự rất quan trọng và cần thực hiện một cách chính xác. Để đảm bảo có thể thay thế hết, ngoài việc sử dụng các tệp sưu tầm được thì cũng cần cập nhật các từ viết tắt này theo bộ dữ liệu đang có.

đc được
ms mới
nx nữa
cx cũng
lm làm
mn mọi người
nma nhưng mà
thík thích
bít biết
j gì
iu yêu
xik xinh
...

Hình 12. Các từ viết tắt.

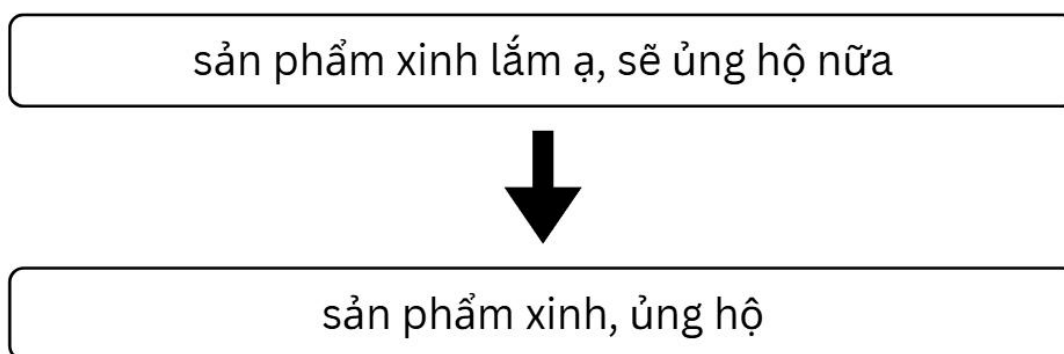
h) Xóa hư từ (stop words)

Stopwords là các từ dùng trong tiếng Việt, thường xuất hiện với tần suất cao trong câu để hỗ trợ cấu trúc ngữ pháp hoặc ngữ điệu, nhưng lại không đóng vai trò quan trọng trong việc biểu đạt nội dung cốt lõi. Một số ví dụ điển hình gồm: "ai", "kìa", "rồi", "cứ",... Do không mang nhiều ý nghĩa về mặt ngữ nghĩa, các stopwords thường được loại bỏ trong quá trình tiền xử lý dữ liệu nhằm tập trung vào các từ khóa chính – những yếu tố quan trọng giúp mô hình phân tích cảm xúc hoạt động hiệu quả hơn.

Cũng giống như các từ viết tắt, để đảm bảo xử lý tốt hơn cần phải xem xét kỹ lưỡng bộ dữ liệu để thêm vào những hư từ đang có.

ai
thì
họ
ấy
bị
chùng
nào
nè
nhiều
cũng
sao
thì
...

Hình 13. Một số hư từ.



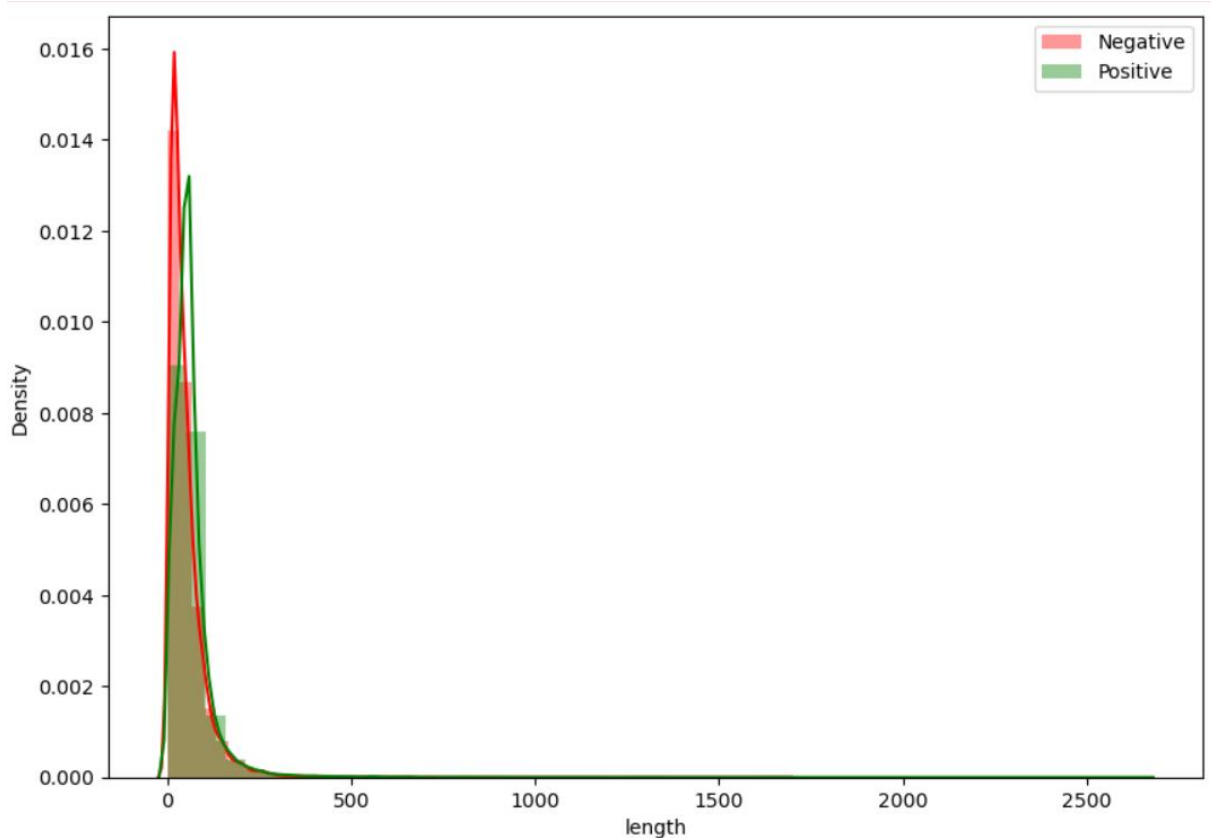
Hình 14. Xử lý hư từ.

i) Xóa các từ không mang tính phân loại

Bên cạnh hư từ, bộ dữ liệu còn có các từ được lặp lại giữa các đánh giá. Vì Shopee cung cấp mẫu đánh giá nên các từ này bị lặp lại. Ví dụ: "màu sắc", "đúng mô tả", "chất liệu", "hình ảnh", "video", "mang tính chất", "nhận xu", "minh họa". Các từ này không mang tính phân loại nên tiến hành xóa khỏi bộ dữ liệu.

j) Outlier

Trong bộ dữ liệu tồn tại những hàng có đánh giá với độ dài lên tới hơn 2500 và điều này rất kì lạ:



Hình 15. Độ dài của đánh giá.

Kiểm tra cho thấy, các đánh giá có độ dài lớn hơn 400 chứa nội dung quảng cáo, không liên quan đến danh mục sản phẩm của đề tài, sẽ gây nhiễu cho việc huấn luyện mô hình nên các đánh giá này cũng được xóa.

Độ dài đánh giá	Đánh giá	Số lượng
400 - 500	hồ ba bể nằm cánh_cung sông gâm địa_hình gồ_ghề cắt xé ngọn núi cao mình mình xen_kẽ t hung_lũng hồ được hình_thành cách hơn triệu năm cuộc kiến_tạo lục_địa đông nam á cuối kỷ đưa một khối nước khổng_lồ diện_tích bề_mặt xấp_xỉ triệu mình chiều dày hơn mình lē n lưng vùng núi đá_vôi tạo hồ ba bể hồ ba b ể nằm độ cao khoảng mình mực nước biển diện _tích mặt_nước hơn được bọc những dãy núi đ á_vôi nhiều hang động những suối ngầm độ sâ u trung_bình hồ mình mùa mưa thể xuống xấp_xỉ mình mùa khô	939
500 - 1000	x le new released collection deadly sins and the oposite thất đại tội tội_lỗi con dễ ph ạm căn nguyên tội_ác ghê_tởm tàn_bạo kinh_h oàng nhất xã_hội loài bản_chất chẳng tồn đi ều tuyệt_đối ngay cả thất đại tội không hoà n_toàn xấu_xa thường phán_xét nhiều hoàn_cả nh thể trở_thành động_lực mạnh_mẽ thúc_đẩy vượt qua chính dần hoàn_thiện hơn kiêu_ngạo cách trân_trọng nỗ_lực bản_thân lười_biếng tạo khoảng nghỉ hồi_phục tham ăn_không bỏ phí những xứng_đáng được thụ_hưởng đồ_ky th úc bản_thân ganh đua_tranh_đấu phần_nộ giải uất_ức khẳng_định quyền_lực tham_lam tạo m ục_tiêu tiến lên sắc_dục khiến tình_yêu sâu _sắc mãnh_liệt khai_thác thất đại tội những biểu_tượng đối_lập nhấn_mạnh thông_điệp ch ế_ngự cấm_dỗ giải_phóng một_cách đúng_mực t iết chế tác hãy xem dưới store hà_nội lê th anh nghị hai trưng hà_nội nguyên trái thanh _xuân hà_nội hotline store hồ chí minh hạt phường quận hồ chí minh the new playground đối_diện lê lai quận hồ chí minh hotline hã y phụ_kiện điểm nhấn cuối_cùng outfit bạn	301
1000 - 1500	nam thư xin_lỗi ba mẹ khẳng_định việc chưa kết_thúc đuổi cùng mới hợp_báo chính_thức m ột thời_gian dài vương lùm xùm mạng xã_hội_ diễn_viên nam thư xúc_động gửi lời xin_lỗi gia_đình bạn_bè nói hôm_nay muốn gửi lời xi	47

	<p>n_lỗi tới cha_mẹ con xin_lỗi ba mẹ suốt một tháng qua im_lặng ba mẹ con dám nói ba mẹ hãy tin đợi con_con biết ba mẹ rất lo_buồn bản_thân con cố_gắng giải_quyết chuyện một_cách minh_bạch rõ_ràng xin_lỗi tất_cả đồng_nghiep vô_tình cuốn câu_chuyện cảm_thấy vô_cùng áy_náy xin_lỗi cả học_trò nhiều học_trò đang chịu nhiều áp_lực xin_lỗi cả đơn_vị chủ_quản quản_lý trực_tiếp nam thư công_ty thiệt_hại thất_thoát nhiều tiền_bạc ảnh_hưởng nhân_viên vụ_việc lùm_xùm gây chấn_động xoay quanh việc tố tiểu tam phá_hoại gia_đình khác nam thư nói xin cảm_ơn đồng_nghiep bạn_bè một tháng im_lặng động_viên dỗi ngày cảm_ơn nhà_báo ngày đầu_tiên chia_sẻ ấm_lòng lắm biết tin chọn tin việc chưa thể kết_thúc được đợi quan điều_tra đưa kết quan điề_tra không_thể làm_việc nhanh gấp được đủ quy_trình làm_việc cấp_độ nghiêm_trọng vụ_việc quan điều_tra làm_việc ngày hai tháng hoặc bốn tháng chắc_chắn công_ty đuổi vụ_việc cùng kết chính_thức công_bố mong_muốn nhiều góc nhìn hơn câu_chuyện</p>	
1500 - 2000	<p>đúng hồng vải dự_án truyện i đọc hiểu điểm đọc văn_bản thực_hiện yêu_cầu trắc_nghiem tiếng trống thu không chòi huyện nhỏ tiếng một vang xa gọi chiều phương tây đỏ_rực lửa cháy những đám mây ánh hồng hòn than tàn dầy tre làng mặt đen cắt hình rõ_rệt nền trời chiều chiều một chiều êm_ả ru văng_vẳng tiếng ếch_nhái kêu ran ngoài đồng_ruộng gió nhẹ đưa cửa_hàng hơi tối muối bắt_đầu vo_ve_liên ngồi yên_lặng thuốc sơn đen đôi mắt bóng_tối ngập đầy dần buồn chiều quê thăm_thía tâm_hồn ngây_thơ liên không hiểu thấy lòng buồn man_mác giờ khắc ngày tàn thấp đèn lên liên nghe tiếng an liên đứng dậy trả_lời hăng thông_thả một lát được ngồi kéo muối an bỏ diêm xuống bàn cùng ngoài chõng ngồi chiếc chõng nan lún xuống kêu cọt_két chõng gãy ừ bảo mẹ mua khác thay hai gượng_nhẹ ngồi yên nhìn phố nhà lên_đèn cả đèn treo nhà phở mỹ đèn hoa_kì leo_lết nhà cửa đèn dây sáng xanh hiệu khách những nguồn ánh_sáng đều chiếu ngoài phố khiến cát lấp_lánh chỗ đường mấp_mô những hòn đá nhỏ một sáng một</p>	15

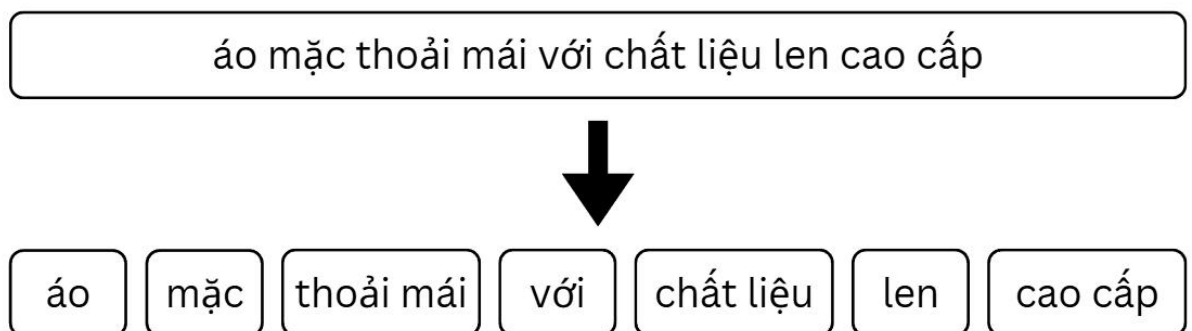
	<p>tối chợ họp giữa phố vẫn lâu hết tiếng ồn ào mất đất rác_rưởi vỏ bưởi vỏ thị lá nhãn bĩa mĩa một mùi âm_âm bốc lên hơi nóng ban_ngày lẫn mùi cát bụi quen_thuộc quá khiến liên_tưởng mùi riêng đất quê_hương một_vài bán hàng muộn đang thu_xếp_hàng định gánh xỏ sẵn quang đứng nói_chuyện ít câu đùa trẻ_con nhà nghèo ven chợ cúi lom_khom mặt_đất đi tìm_tòi nhặt_nhặt thanh nửa thanh tre bắt_thể dùng được bán hàng liên trông thấy động_lòng thương chính không tiền hãy viết bài văn khoảng chữ nêu cảm_nhận bức tranh phố huyện chiều tàn nhìn bé liên tác_phẩm hai đứa trẻ thạch lam qua đoạn trích</p>	
<p>2000 - 2500</p>	<p>lưu quan_trọng đi thi tốt_nghiệp đi thi ăn_uống nhẹ_nhàng bụng đói uống sữa ăn_uống những thứ lạ ngoài đường đi vệ_sinh phòng thi địa_điểm thi phút dụng_cụ đi thi bút_bi cùng loại cùng màu mực xanh hoặc đen nên chuẩn_khoảng cây bút_chì loại bạn nên chuốt nhọn chuốt mài vệt đầu tô nhanh công_đoạn chuốt mài chuẩn nhà chuẩn khoảng cây gôm loại xịn nên mua cục to để xoá máy_tính mang bấy nhớ kiểm_tra pin lâu quá chưa thay pin mang thay pin những chỗ thay pin đồng_hồ đều thay pin máy_tính tránh trường_hợp vô thi vỗ bôm_bốp máy_tính phiếu dự thi chứng_minh_thư đồng_hồ nên mang đồng_hồ canh giờ chai nước nhỏ bóc nhãn nên pha nước trà đường mang nhấm_nháp nhấm_nháp uống nhiều đi tất_cả giấy_tờ bút viết máy_tính hết túi đựng tài_liệu clear bag đề_phòng nhốn_nháo hoặc gấp giang_hồ hiểm_ác rất dễ hack mất đồ làm hoi_hộp ngồi thẳng lưng hít thở sâu uống ngụm nước nhỏ nhận được đề kiểm_tra đề mã_đề số trang chất_lượng in điền tô đúng thông_tin phiếu trả_lời trắc_nghiệm chiến_thuật làm bài đọc đề suy_nghi hướng đơn_giản nhất để khó mặc đề thi được xếp dễ khó điều_tương_đối cố_gắng câu_đầu nháp cẩn_thận chia vùng rõ_ràng nháp xong câu cách đoạn nháp câu tiếp giấy nháp được xin tô đáp_án làm xong câu tô câu tránh tính_trạng cuối giờ cuống câu tô đáp_án hoặc làm anh tô bạn do lộn làm cấn_thận đầu tránh suy_nghi kiểu làm nhanh lát dò hạn_chế trao_đổi bắt_chuyện thi xong t</p>	<p>4</p>

	ý_lệ chọi cao nên đừng tin đũa hết canh thờ i_gian cách lụi câu cuối lụi hên xui hên xu i xởi_lởi đừng toan_tính tất_cả do ăn tính bằng trời tính nên lụi đáp_án hiện lên đầu đầu_tiên chọn suy_nghĩ dành thời_gian làm c âu_đầu nhớ đừng nhàu tờ phiếu trả_lời trắc_ nghiệm không vẽ bậy lên giấy nháp quy_chế đ ình vi_phạm thi một nhà đừng mở bất_kì tran g mạng tìm đáp_án không do bộ giáo_dục cung _cấp tránh hoang_mang giữ tinh_thần thi tiế p quan_trọng chuẩn một lượng kiến_thức đầy_ đủ biết địch biết trăm trận trăm thắng nên hãy trang hành_trang tốt ôn_luyện kỹ nhớ gi ữ một đầu lạnh cuối_cùng chuẩn một tâm_hồn đẹp chúc tất_cả sĩ tử đều thi tốt cố_gắng v ững_vàng tâm_lý_chiến_thắng không er nhớ no te bài đi thi chuẩn tốt nhất	
--	--	--

Bảng 3. Đánh giá có độ dài > 400.

k) Tách từ

Trong tiếng Việt, dấu cách không dùng để phân tách từ mà để ngắt các âm tiết, do đó bài toán tách từ (word segmentation) đóng vai trò đặc biệt quan trọng trong xử lý ngôn ngữ tự nhiên [7]. Việc tách từ không chính xác có thể làm sai lệch ngữ nghĩa của câu, từ đó ảnh hưởng tiêu cực đến hiệu suất và độ chính xác của mô hình học máy. Để giải quyết vấn đề này, đề tài đã lựa chọn sử dụng công cụ ViTokenizer thuộc thư viện **pvi**, một giải pháp hiệu quả giúp thực hiện việc tách từ trong tiếng Việt một cách nhanh chóng và chính xác.



Hình 16. Tách từ.

3.2.2.2. Gán nhãn và cân bằng tập dữ liệu

a) Gán nhãn tập dữ liệu

Rating	Số lượng
5	178566
4	13004
3	12481
2	6059
1	10881

Bảng 4. Số lượng bình luận ở mỗi rating.

Trước khi tiến hành huấn luyện mô hình, cần gán nhãn dữ liệu bằng cách phân loại cảm xúc dựa trên điểm đánh giá (Rating) từ khách hàng. Dựa vào Bảng 3-2, dữ liệu thu thập được chia thành hai nhóm theo quy tắc sau:

- **Rate ≤ 3** : Các bình luận có điểm số từ 3 trở xuống được gán nhãn tiêu cực (Negative).
- **Rate > 3** : Các bình luận có điểm trên 3 được gán nhãn tích cực (Positive).

Nhãn	Số lượng
Positive	191570
Negative	29622

Bảng 5. Số lượng bình luận ở mỗi nhãn

b) Cân bằng tập dữ liệu

Có thể nhận thấy sự chênh lệch đáng kể về số lượng giữa hai nhóm dữ liệu tích cực (positive) và tiêu cực (negative), do đó, tập dữ liệu huấn luyện sẽ bằng $0.8 * \min(\text{size}(\text{positive}), \text{size}(\text{negative})) * 2$ nhằm đảm bảo tính cân bằng và cải thiện hiệu quả của mô hình học máy.

Label	Số lượng
0	15533
1	15533

Bảng 6. Số đánh giá sau cân bằng ở mỗi nhãn

3.2.2.3. Xây dựng và đánh giá mô hình

Tập dữ liệu sau bước tiền xử lý gồm 31066 hàng dữ liệu với 15533 hàng dữ liệu mỗi nhãn và 4 cột: comment, nomalized_comment, emoji, label.

a) Xây dựng Emoji sentiment model

* Chuẩn bị dữ liệu

Theo mặc định, các đối tượng vectorizer trong thư viện **scikit-learn** sẽ loại bỏ toàn bộ các ký tự dấu câu (*punctuation*) khỏi văn bản đầu vào trong quá trình tiền xử lý. Do đó, các biểu tượng cảm xúc (*emoji*), vốn thường được biểu diễn dưới dạng các ký tự đặc biệt, cũng sẽ bị loại bỏ khi văn bản được chuyển đổi thành vector. Để khắc phục vấn đề này và giữ lại thông tin cảm xúc quan trọng mà các emoji mang lại, thay vì giữ nguyên biểu diễn ký tự, ta sẽ chuyển đổi chúng sang dạng văn bản thông qua quá trình *decode*. Cụ thể, một đặc trưng mới có tên là emoji_decode sẽ được tạo ra để lưu trữ nội dung biểu tượng cảm xúc dưới dạng văn bản đã giải mã.

	comment	nomalized_comment	emoji_decode	label
0	Màu áo ko giống, chất mềm mát nhưng ...	màu áo giống chất mềm mát khá mỏng ...	face_holding_back_tears	0
1	Màu sắc:trang\n\nỂ nó ngắn hơn tui tư...	trang ề ngắn hơn tôi tưởng bánh muố...	smiling_face_with_tear	0
2	Thấy đánh giá bên trong vải lưới mà ...	thấy giá vải lưới vải thất_vọng mậ...	smiling_face_with_tear	0
3	💎💎 Gót quai đá đẹp xuất sắc fullbox\n👍...	gót quai đá đẹp xuất_sắc_màu trắng_đ...	money_with_wings sun_with_face sparkles sparkl...	0
4	Giao hàng nhanh, đóng gói ổn. Tuy nhiên ...	giao hàng nhanh đóng_gói ổn nhiên màu x...	smiling_face_with_tear	0

Hình 18. Chuẩn bị dữ liệu cho Emoji sentiment model.

Tiến hành chuẩn bị tập dữ liệu huấn luyện (*training data*) và tập dữ liệu kiểm tra (*test data*) cho mô hình phân tích cảm xúc bình luận. Cụ thể, tập dữ liệu sau khi được tiền xử lý sẽ được chia thành hai phần: 70% dùng để huấn luyện mô hình và 30% còn lại dùng để đánh giá hiệu suất mô hình.

* Huấn luyện mô hình

Tiến hành huấn luyện mô hình áp dụng cross-validation chia dữ liệu huấn luyện thành 10 phần. Kết quả:

Classifiers	Feature Extraction Techniques			
	Bag-of-Words	TF-IDF	Bag-of-Words (min_df=2)	TF-IDF (min_df=2)
SVM	82.13	82.95	82.18	82.27
Logistic Regression	82.72	82.84	81.65	82.07
Random Forest	87.13	87.3	86.23	84.42
XGBoost	74.78	74.59	74.78	75.03
K-NN	75.09	75.35	74.46	77.44
Bernoulli	83.87	83.87	82.01	82.01

Bảng 7. Kết quả so sánh các mô hình dựa vào độ chính xác trên tập huấn luyện.

Classifiers	Feature Extraction Techniques			
	Bag-of-Words	TF-IDF	Bag-of-Words (min_df=2)	TF-IDF (min_df=2)
SVM	78.52	79.17	77.7	78.65
Logistic Regression	77.77	79.35	77.77	79.35
Random Forest	76.81	77.25	77.06	77.38
XGBoost	74.78	74.59	74.78	75.03
K-NN	75.09	75.35	74.46	77.44
Bernoulli	79.16	79.18	79.03	79.03

Bảng 8. Kết quả so sánh các mô hình dựa vào độ chính xác trên tập kiểm tra.

Từ số liệu so sánh giữa các phương pháp trích xuất đặc trưng và các mô hình học máy (Bảng 3-4), phương pháp TF-IDF hoạt động tốt nhất đối với 3 mô hình SVM, Logistic Regression và Bernoulli với độ chính xác trung bình lần lượt là 85.95%, 85.84% và 83.87%. Tiến hành train cross-validation và lựa ra các hyperparameter set cho từng

model. Sau đó, train lại từng model với hyperparameter set trả về trên toàn bộ training data (không chia validation).

*** Đánh giá mô hình**

Mô hình	Train accuracy	Test accuracy	Train roc auc	Test roc auc
SVC - [kernel: rbf]	83.61	78.67	83.18	78.19
Logistic Regression - [solver: liblinear]	83.67	78.37	83.25	77.85
Bernoulli	84.18	84.09	84.09	77.53

Bảng 9. Đánh giá các mô hình Emoji sentiment comment.

Hai model Logistic và SVC có hiệu suất có thể chấp nhận được khi độ lệch giữa `train_acc` và `test_acc` dưới 6%.

Tuy nhiên, lúc này ta có thể thấy test data của ta có sự nổi trội hơn của một trong hai class positive và negative so với class còn lại dựa vào `test_roc_auc`. Điều này có thể do hai lí do, test data của ta thực sự bị lệch hoặc model của ta bị overfitting, nhưng khả năng cao là test data bị lệch vì `test_roc_auc` có hiệu suất vẫn chấp nhận được.

b) Xây dựng Comment sentiment model

* Chuẩn bị dữ liệu

Tiến hành chuẩn bị tập dữ liệu huấn luyện (*training data*) và tập dữ liệu kiểm tra (*test data*) cho mô hình phân tích cảm xúc bình luận. Cụ thể, tập dữ liệu sau khi được tiền xử lý sẽ được chia thành hai phần: 70% dùng để huấn luyện mô hình và 30% còn lại dùng để đánh giá hiệu suất mô hình.

* Huấn luyện mô hình

Tiến hành huấn luyện mô hình áp dụng cross-validation chia dữ liệu huấn luyện thành 10 phần. Kết quả (top 6 model cho kết quả tốt nhất):

Classifiers	Feature Extraction Techniques			
	TF-IDF	TF-IDF (min_df=5)	TF-IDF (min_df=10)	TF-IDF (min_df=20)
SVC - [kernel: rbf]	94.82	94.37	94.12	93.7
SVC - [kernel: linear]	85.85	88.28	87.76	85.68
SVC - [kernel: poly]	97.02	97.13	96.96	96.66
Logistic Regression - [solver: newton-cg]	88.58	89.58	87.2	85.63
Logistic Regression - [solver: lbfgs]	88.22	87.58	87.2	85.64
Logistic Regression - [solver: liblinear]	88.72	89.58	87.2	85.55

Bảng 10. Kết quả so sánh top 6 mô hình dựa vào độ chính xác trên tập huấn luyện

Classifiers	Feature Extraction Techniques			
	TF-IDF	TF-IDF (min_df=5)	TF-IDF (min_df=10)	TF-IDF (min_df=20)
SVC - [kernel: rbf]	86.87	86.88	86.81	86.74
SVC - [kernel: linear]	85.85	85.87	85.7	85.68
SVC - [kernel: poly]	86.07	86.08	85.95	85.82
Logistic Regression - [solver: newton-cg]	85.71	85.74	85.65	85.63
Logistic Regression - [solver: lbfgs]	85.71	85.74	85.7	85.64
Logistic Regression - [solver: liblinear]	85.71	85.75	85.69	85.55

Bảng 11. Kết quả so sánh top 6 mô hình dựa vào độ chính xác trên tập kiểm tra.

Từ số liệu so sánh giữa các phương pháp trích xuất đặc trưng và các mô hình học máy (Bảng 3-4), mô hình SVC kernel rbf cho kết quả tốt nhất với train accuracy 94,37% và 86.88% cho test accuracy; mô hình SVC kernel poly có hiện tượng overfitting với mọi vectorizer. Tiếp theo, tiến hành train cross-validation và lựa chọn các hyperparameter set cho từng model. Sau đó, train lại từng model với hyperparameter set trả về trên toàn bộ training data (không chia validation).

Hình 20. Train cross-validation và lựa chọn hyperparameter set cho từng model.

* Đánh giá mô hình

Mô hình SVC kernel rbf vẫn đạt hiệu suất tốt nhất trên tập kiểm tra mà không

Mô hình	Train accuracy	Test accuracy	Train roc auc	Test roc auc
SVC - [kernel: rbf]	94.24	86.64	94.24	86.65
Logistic Regression - [solver: liblinear]	87.57	86.06	87.57	86.07
Logistic Regression - [solver: newton-cg]	87.57	86.06	87.57	86.07
Logistic Regression - [solver: lbfgs]	87.57	86.06	87.57	86.07
SVC - [kernel: linear]	88.31	86.02	88.31	86.03
SVC - [kernel: poly]	97.06	85.87	97.06	85.89

bị overfitting nghiêm trọng. Với mô hình SVC kernel linear mặc dù đơn giản hơn nhưng vẫn có hiệu suất tương đương và khả năng diễn giải tốt hơn. Các mô hình Logistic Regression có sự tăng nhẹ test accuracy, và cả ba solver đều có kết quả tương đương, tuy nhiên thời gian training của solver liblinear là nhỏ nhất (0.59s) nên sẽ ưu tiên giữ lại mô hình này.

c) Xây dựng Comment sentiment model bằng LSTM

* Chuẩn bị dữ liệu

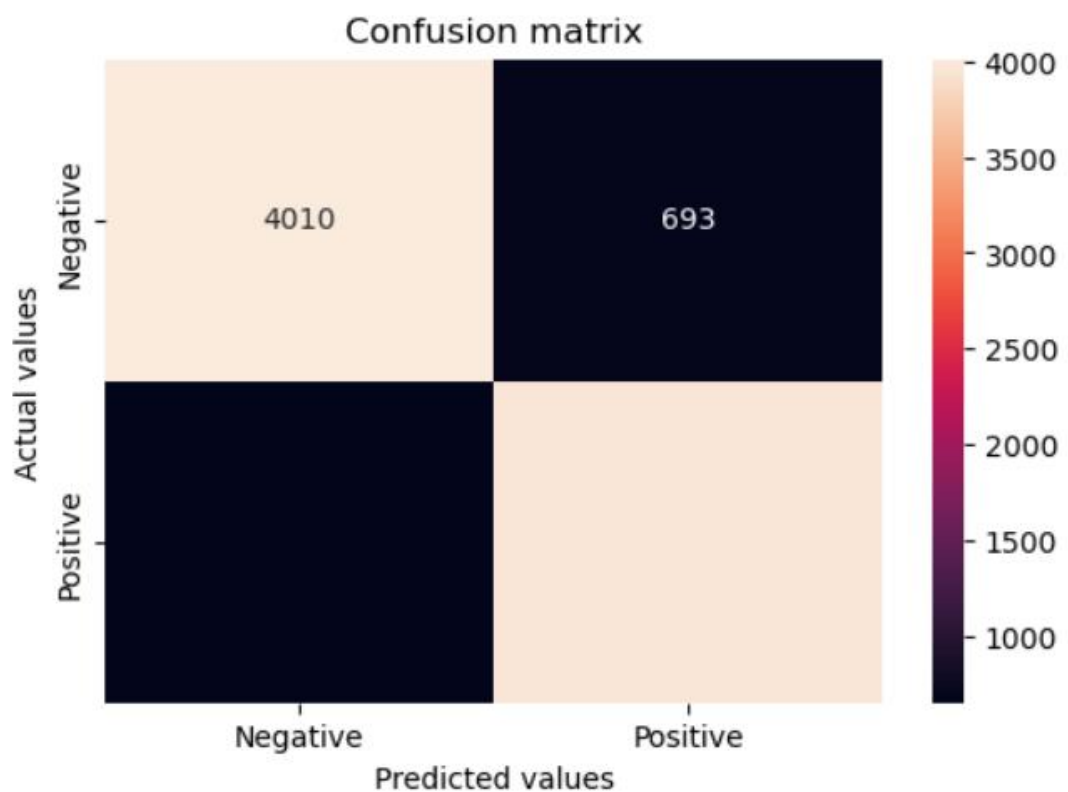
Dữ liệu huấn luyện cho mô hình này là các đánh giá đã qua tiền xử lý ở cột “nomalized_comment”. Tiến hành chuẩn bị tập dữ liệu huấn luyện (*training data*) và tập dữ liệu kiểm tra (*test data*) cho mô hình phân tích cảm xúc bình luận. Cụ thể, tập dữ liệu sau khi được tiền xử lý sẽ được chia thành hai phần: 70% dùng để huấn luyện mô hình và 30% còn lại dùng để đánh giá hiệu suất mô hình.

* Huấn luyện mô hình và đánh giá mô hình

Tiến hành huấn luyện mô hình với hyper-parameter “num_words” = None để Tokenizer (Tensorflow cung cấp) tự quyết định. Bên cạnh đó, sử dụng cơ chế kiểm soát ModelCheckpoint của TensorFlow để lưu lại bộ trọng số (weights) tốt nhất của mô hình. Kết quả:

Thuật toán	LSTM	
	Positive	Negative
Accuracy	86	
F1_score	85	86
Recall	86	85
Precision	85	86

Bảng 13. Kết quả huấn luyện mô hình LSTM lần 1.



Hình 21. Ma trận nhầm lẫn của mô hình LTSM lần 1.

Có thể nhận thấy rằng mô hình LSTM đạt độ chính xác tương đối cao, vượt mức 86%. Số lượng sai lệch (FP và FN) gần tương đương (693 và 655), cho thấy mô hình không bị lệch về một loại sai nào. Tỷ lệ lỗi thấp hơn 15% ở mỗi lớp, phù hợp với độ chính xác đã nêu. Đáng chú ý, mặc dù các siêu tham số (hyper-parameters) được thiết lập tự động bởi TensorFlow, mô hình vẫn cho thấy khả năng dự đoán cân đối giữa hai lớp "negative" và "positive". Nhờ vào sự cân bằng này, việc đánh giá thêm bằng chỉ số ROC-AUC là không thực sự cần thiết trong trường hợp này.

d) Kết hợp mô hình và đánh giá

Đề tài đã huấn luyện được 3 mô hình Emoji sentiment và 5 mô hình Comment sentiment. Vậy khi kết hợp sẽ có 15 mô hình. Sau khi kết hợp tiến hành đánh giá dựa trên Accuracy. Kết quả:

Mô hình	Accuracy
logistic_svc_linear	87.85
logistic_logistic	88.15
Logistic_svc_rbf	91.11
Logistic_lstm_1	91.26
Logistic_lstm_uw2	89.63
Svc_svc_linear	87.26
Svc_logistic	86.37
Svc_svc_rbf	88.89
Svc_lstm_1	91.11
Svc_lstm_uw_2	89.33
Bernoulli_svc_linear	86.67
Bernoulli_logistic	86.52
Bernoulli_svc_rbf	88.89
Bernoulli_lstm_1	90.96
Bernoulli_lstm_uw2	88.89

Bảng 14. Độ chính xác của các mô hình kết hợp.

Sau khi tổng hợp kết quả, có tổng cộng 15 mô hình đạt độ chính xác (accuracy) trên 86%. Trong số đó, 4 mô hình đạt độ chính xác vượt mốc 90%, và 3 mô hình trong nhóm này sử dụng kiến trúc LSTM cho bài toán phân tích cảm xúc bình luận. Với kết quả tốt nhất, đề tài sẽ sử dụng mô hình Logistic emoji sentiment kết hợp với mô hình LSTM_1 để xây dựng hệ thống dự đoán cảm xúc.

3.2.3. Mô hình dự đoán:

3.2.3.1. Xử lý dữ liệu đưa vào

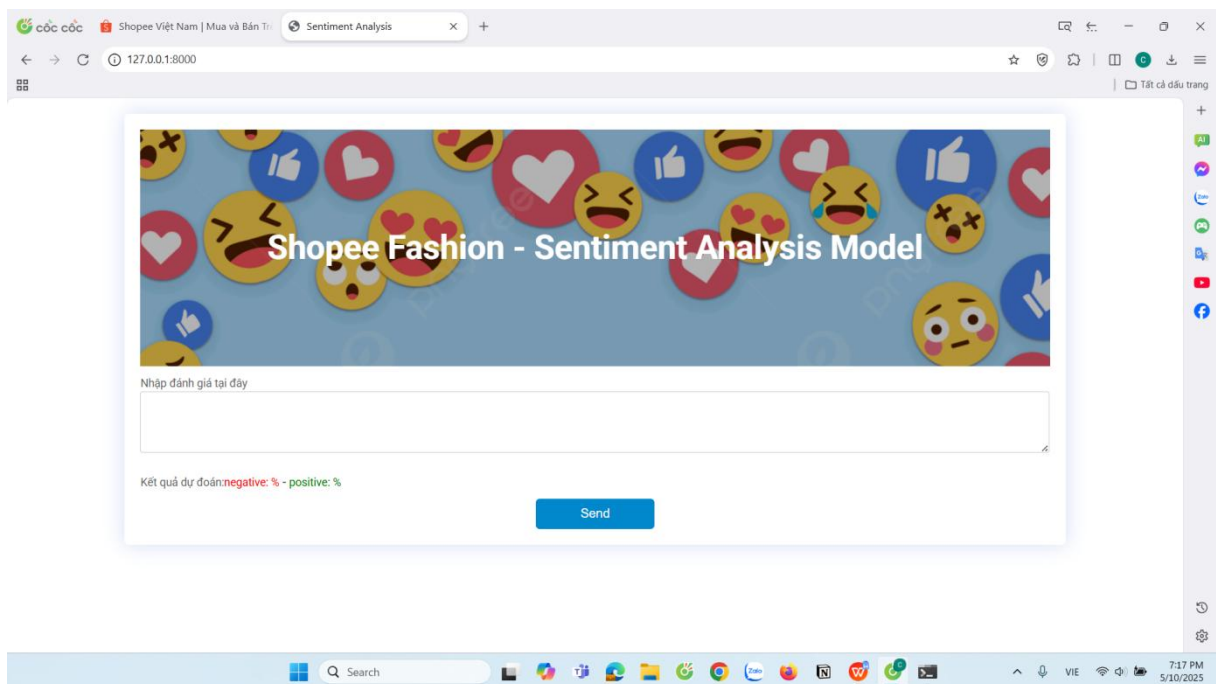
Ở bước này, mục tiêu trọng tâm là làm sạch và chuẩn bị dữ liệu một cách kỹ lưỡng nhằm đảm bảo chất lượng đầu vào cho các bước dự đoán và phân tích tiếp theo. Các thao tác tiền xử lý được thực hiện tương tự như quy trình xử lý dữ liệu ở bước xây dựng mô hình huấn luyện.

3.2.3.2. Dự đoán cảm xúc

Hệ thống sẽ tiến hành đưa dữ liệu đã xử lý vào mô hình kết hợp được chọn để tiến hành dự đoán cảm xúc khách hàng của đánh giá được nhập.

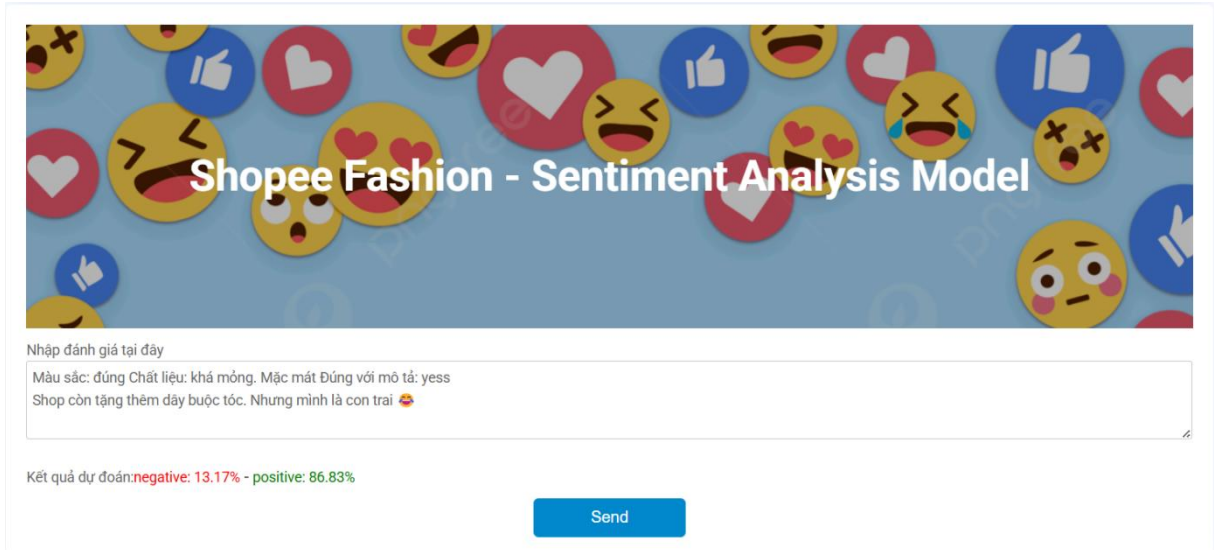
3.3. Kết quả đạt được

3.3.1. Chức năng



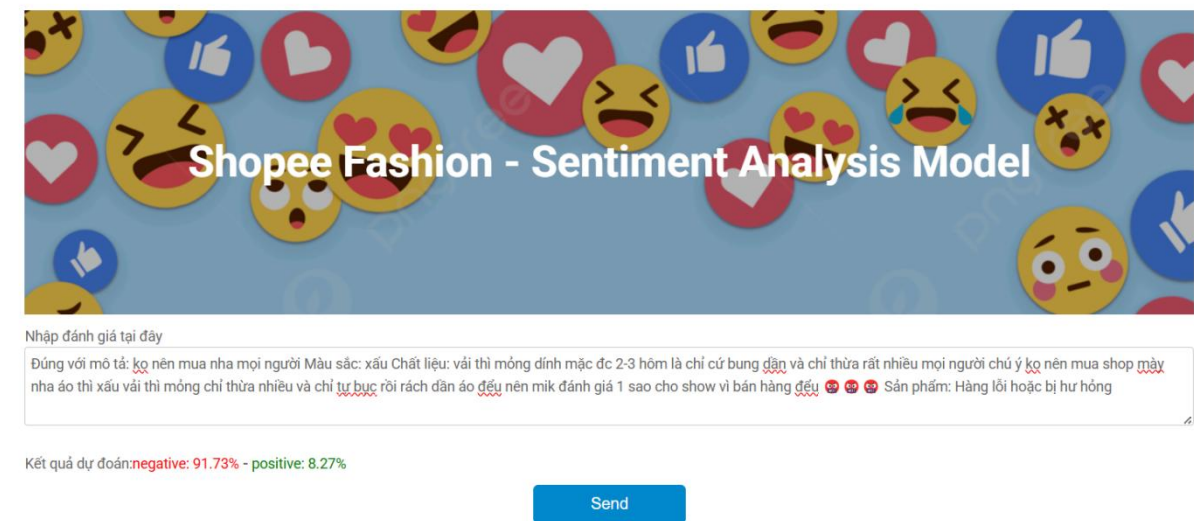
Hình 22. Trang chủ hệ thống dự đoán cảm xúc khách hàng.

Trang chủ cho phép người dùng nhập đánh giá. Nhấn “Send” để hệ thống phân tích và đưa ra xác suất phần trăm tích cực và tiêu cực của đánh giá.



The screenshot shows the main interface of the 'Shopee Fashion - Sentiment Analysis Model'. The header features a blue background with various colorful emojis. Below the header, there is a text input field with the placeholder 'Nhập đánh giá tại đây'. The input field contains the text: 'Màu sắc: đúng Chất liệu: khá mỏng. Mặc mát Đúng với mô tả: yess Shop còn tặng thêm dây buộc tóc. Nhưng mình là con trai 🤔'. Below the input field, the system's output is displayed: 'Kết quả dự đoán: negative: 13.17% - positive: 86.83%'. A blue 'Send' button is located at the bottom right of the interface.

Hình 23. Hệ thống trả kết quả với đánh giá 5 sao.



The screenshot shows the same interface as Figure 22, but with a different input and output. The input text is: 'Đúng với mô tả: kô nên mua nha mọi người Màu sắc: xấu Chất liệu: vải thì mỏng dính mặc đc 2-3 hôm là chỉ cứ bung dần và chỉ thừa rất nhiều mọi người chú ý kô nên mua shop này nha áo thì xấu vải thì mỏng chỉ thừa nhiều và chỉ tự bực rồi rách dần áo đều nên mik đánh giá 1 sao cho show vì bán hàng đều 🤔 🤔 🤔 Sản phẩm: Hàng lỗi hoặc bị hư hỏng'. The output shows a high negative sentiment: 'Kết quả dự đoán: negative: 91.73% - positive: 8.27%'. The 'Send' button remains at the bottom right.

Hình 24. Hệ thống trả kết quả với đánh giá 1 sao.

3.3.2. Kết quả mô hình

Tất cả các mô hình có thành phần Comment Sentiment Model sử dụng kiến trúc lstm_1 đều đạt độ chính xác cao. Trong khi đó, các mô hình sử dụng các thuật toán phân loại truyền thống thường dự đoán sai nhiều ở các đánh giá mang cảm xúc tích cực, dễ bị nhầm lẫn sang tiêu cực.

Từ kết quả này, có thể rút ra nhận định rằng đối với tập dữ liệu đánh giá sản phẩm trên Shopee, riêng phần nội dung bình luận, các mô hình Deep Learning – đặc biệt là LSTM – cho hiệu quả vượt trội hơn hẳn. Điều này có thể bắt nguồn từ việc cấu trúc và ngữ cảnh trong comment ảnh hưởng đáng kể đến khả năng phân loại cảm xúc.

Trong số 15 mô hình được thử nghiệm, những mô hình có phần Comment Sentiment Model tích hợp lstm_1 cho kết quả khả quan nhất, do đó có thể được ưu tiên giữ lại để cải tiến hoặc phát triển thêm trong các nghiên cứu và ứng dụng sau này.

Chương 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

4.1. Kết luận

Qua quá trình nghiên cứu và phát triển, đề tài đã đạt được những kết quả sau:

Xây dựng thành công 15 mô hình dự đoán cảm xúc khách hàng với độ chính xác khá cao (trên 86%), trong đó có 4 mô hình đạt độ chính xác trên 90%.

Bước đầu xây dựng hệ thống dự đoán cảm xúc.

4.2. Hướng phát triển

Mặc dù đề tài đã đạt được những kết quả khả quan, vẫn còn nhiều tiềm năng để tiếp tục phát triển và hoàn thiện. Một số định hướng mở rộng trong tương lai có thể bao gồm:

Nâng cao độ chính xác của mô hình: Tích hợp các kiến trúc học sâu tiên tiến hơn như BiLSTM, GRU, hoặc transformer (BERT, RoBERTa) nhằm tăng khả năng hiểu ngữ cảnh và sắc thái cảm xúc trong bình luận người dùng.

Mở rộng và đa dạng hóa tập dữ liệu huấn luyện: Thu thập thêm dữ liệu từ nhiều sàn thương mại điện tử khác nhau (như Lazada, Tiki, Amazon) để tăng tính khái quát và khả năng thích ứng của mô hình với nhiều loại ngữ cảnh đánh giá khác nhau.

Tối ưu hóa thời gian xử lý và phản hồi kết quả: Tăng hiệu suất xử lý để đảm bảo hệ thống có thể phân tích và trả kết quả nhanh chóng, phù hợp với các ứng dụng thời gian thực như chatbot tư vấn hoặc phân tích xu hướng đánh giá tức thời.

Bổ sung chức năng hỗ trợ người dùng và doanh nghiệp: Bao gồm gợi ý cải thiện sản phẩm dựa trên phân tích cảm xúc tiêu cực, hoặc dashboard trực quan giúp doanh nghiệp nắm bắt xu hướng phản hồi của khách hàng theo thời gian.

Cải tiến giao diện người dùng: Thiết kế hệ thống đơn giản, trực quan, dễ sử dụng cho cả người dùng cuối lẫn nhà quản trị, kèm theo các hướng dẫn sử dụng chi tiết và tính năng hỗ trợ kỹ thuật nhanh chóng.

Tích hợp tính năng thu thập dữ liệu tự động: Xây dựng công cụ crawler để tự động cập nhật và đồng bộ dữ liệu bình luận mới từ các nền tảng thương mại điện tử vào cơ sở dữ liệu, phục vụ cho quá trình huấn luyện và đánh giá mô hình liên tục.

Việc phát triển hệ thống phân tích cảm xúc không chỉ giúp doanh nghiệp nắm bắt nhanh chóng và chính xác quan điểm của khách hàng mà còn góp phần cải thiện chất lượng dịch vụ, nâng cao trải nghiệm người dùng, đồng thời tạo nền tảng dữ liệu đáng tin cậy cho việc ra quyết định chiến lược trong kinh doanh.

TÀI LIỆU THAM KHẢO

- [1] M. Wankhade, A. C. S. Rao and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, p. 5731–5780, 2022.
- [2] B. Liu, *Sentiment Analysis and Opinion Mining*, New York: Morgan & Claypool, 2012.
- [3] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, 2014.
- [4] S. Das, A. Dey, A. Pal and N. Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," *International Journal of Computer Applications*, vol. 115, no. 9, pp. 31-41, 2015.
- [5] M. Loukili, F. Messaoudi and M. E. Ghazi, "Sentiment Analysis of Product Reviews for E- Commerce Recommendation based on Machine Learning," *International Journal of Advances in Soft Computing and its Applications*, vol. 15, no. 1, pp. 1-13, 2023.
- [6] R. Ahuja, A. Chug, S. Kohli, S. Gupta and P. Ahuja, "The Impact of Features Extraction on the Sentiment Analysis," *Procedia Computer Science*, vol. 152, pp. 341-348, 2019.
- [7] H. H. Thien, T. C. L. Daphne, T. T. Kiet and T. H. Vinh, "Sentiment Analysis based on word vector representation for short comments in Vietnamese language," *NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 165-169, 2022.
- [8] H. D. Abubakar, M. Umar and M. A. Bakale³, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec,"

Sule Lamido University Journal of Science & Technology, vol. 4, no. 1-2, pp. 27-33, 2022.

- [9] W. Ramadhan, S. A. Novianty and S. C. Setianingsih, "Sentiment analysis using multinomial logistic regression," *International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*, pp. 46-49, 2017.
- [10] N. Đ. L. Bằng, N. V. Hồ and H. T. Thành, "Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực," *Tạp chí Khoa học Đại học Mở Thành phố Hồ Chí Minh*, vol. 16, no. 1, pp. 64-78, 2020.
- [11] J. Xu, "Decoding sentiment: A sentiment analysis model for movie reviews," *International Conference on Machine Learning and Automation*, pp. 31-37, 2023.
- [12] V. Vovk, "The Fundamental Nature of the Log Loss Function," *Lecture Notes in Computer Science*, 2015.
- [13] S. Ding, Z. Zhuu and X. Zhang, "An overview on semi-supervised support vector machine," *Neural Computing and Applications*, vol. 28, no. 5, pp. 969-978, 2017.
- [14] X. Lin, "Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing," in *ISBDAI '20: 2020 2nd International Conference on Big Data and Artificial Intelligence*, 2020.
- [15] T. K. Phung, N. A. Te and T. T. T. Ha, "A machine learning approach for opinion mining online customer reviews," *ACIS International Winter Conference on Software Engineering*, pp. 243-246, 2021.
- [16] B. Liu, "Many Facets of Sentiment Analysis. A Practical Guide to Sentiment Analysis," *Socio-Affective Computing*, pp. 11-39, 2017.

PHỤ LỤC