BỘ GIÁO DỤC VÀ ĐÀO TẠO **BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG**
TRƯỜNG ĐẠI HỌC THỦY LỢI

NGUYỄN ĐỨC ANH

Phân tích cảm xúc của khách hàng trong các bình luận sản phẩm trên các sàn thương mại điện tử

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2026

BỘ GIÁO DỤC VÀ ĐÀO TẠO **BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG**
TRƯỜNG ĐẠI HỌC THỦY LỢI

NGUYỄN ĐỨC ANH

Phân tích cảm xúc của khách hàng trong các bình luận sản phẩm trên các sàn thương mại điện tử

Ngành: Công nghệ thông tin

Mã số:

NGƯỜI HƯỚNG DẪN: TS.Nguyễn Mạnh Hiển

HÀ NỘI, NĂM 2026

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
---------★---------

NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Họ tên sinh viên: NGUYỄN ĐỨC ANH
Hệ đào tạo: Đại học chính quy
Lớp: 63CNTT4
Ngành: Công nghệ thông tin
Khoa: Công nghệ thông tin

1- TÊN ĐỀ TÀI: PHÂN TÍCH CẢM XÚC KHÁCH HÀNG TRONG CÁC BÌNH LUẬN SẢN PHẨM TRÊN CÁC SÀN THƯƠNG MẠI ĐIỆN TỬ

2- NỘI DUNG CÁC PHẦN THUYẾT MINH VÀ TÍNH TOÁN: Tỷ lệ %

Nội dung các phần	Tỷ lệ %
Chương 1: Tổng quan về bài toán phân tích cảm xúc trong bình luận TMĐT	10%
Chương 2: Cơ sở lý thuyết	35%
Chương 3: Phân tích và thiết kế hệ thống	20%
Chương 4: Thực nghiệm và đánh giá	35%

3. GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN

Phần	Họ và tên giáo viên hướng dẫn
Chương 1: Tổng quan về bài toán phân tích cảm xúc trong bình luận TMĐT	TS. Nguyễn Mạnh Hiển
Chương 2: Cơ sở lý thuyết	
Chương 3: Phân tích và thiết kế hệ thống	
Chương 4: Thực nghiệm và đánh giá	

4. NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Ngày ..... tháng ..... năm 2025

Trưởng Bộ môn
(Ký và ghi rõ Họ tên)

Giáo viên hướng dẫn chính
(Ký và ghi rõ Họ tên)

Nhiệm vụ Đồ án tốt nghiệp đã được Hội đồng thi tốt nghiệp của Khoa thông qua.

Ngày. . . . .tháng. . . . .năm 2025

Chủ tịch Hội đồng

(Ký và ghi rõ Họ tên)

Sinh viên đã hoàn thành và nộp bản Đồ án tốt nghiệp cho Hội đồng thi ngày..... tháng... năm 2026

Sinh viên làm Đồ án tốt nghiệp
(Ký và ghi rõ Họ tên)

TRƯỜNG ĐẠI HỌC THUỶ LỢI
KHOA CÔNG NGHỆ THÔNG TIN

BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP

Tên đề tài: Phân tích cảm xúc của khách hàng trong các bình luận sản phẩm trên các sàn thương mại điện tử

Sinh viên thực hiện: Nguyễn Đức Anh

Lớp: 63CNTT4

Mã sinh viên: 2151062703

Số điện thoại: 0978898633

Email: 2151062703@e.tlu.edu.vn

Giảng viên hướng dẫn: TS. Nguyễn Mạnh Hiển

TÓM TẮT ĐỀ TÀI

Đề tài nghiên cứu phương pháp và xây dựng hệ thống phân tích cảm xúc (Sentiment Analysis) cho các bình luận sản phẩm trên các sàn thương mại điện tử (ví dụ: Shoppee, Lazada, Tiki). Mục tiêu của đề tài là thu thập dữ liệu đánh giá/bình luận của người dùng, thực hiện tiền xử lý văn bản tiếng việt, xây dựng và so sánh các mô hình phân loại cảm xúc(nhị phân, tam phân hoặc đa lớp), đánh giá hiệu năng mô hình bằng các chỉ số phù hợp, đồng thời triển khai một prototype giúp doanh nghiệp theo dõi mức độ hài long khách hàng theo sản phâm và theo thời gian. Đề tài kết hợp các kỹ thuật xử lý ngôn ngữ tự nhiên(NLP), học máy (Machine Learning) và thực tiễn triển khai phần mềm.

Nhận diện cảm xúc trong văn bản là một trong những bài toán quan trọng trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên (NLP), đặc biệt trong bối cảnh dữ liệu văn bản ngày càng gia tăng từ các nền tảng mạng xã hội và phương tiện truyền thông số. Đề tài này tập trung nghiên cứu và phát triển mô hình học máy để nhận diện cảm xúc trong văn bản tiếng

Việt, dễ mở rộng và hiệu quả cao đối với dữ liệu văn bản ngắn như bình luận thương mại điện tử.

Cụ thể bình luận tiếng Việt sau khi được làm sạch và chuẩn hóa sẽ được biểu diễn dưới dạng đặc trưng số bằng các phương pháp biểu diễn văn bản phổ biến như Bag-of-Words(BoW) và TF-IDF (Term Frequency-Inverse Document Frequency), có thể kết hợp n-gram để khai thác các cụm từ quan trọng trong cảm xúc(ví dụ: "không tốt", "rất hài lòng"," "giao hàng chậm"). Trên cơ sở đặc trưng này, đề tài sử dụng mô hình Support Vector Machine(SVM) để phân loại văn bản nhờ khả năng xử lý tốt dữ liệu thưa và có số chiều lớn (đặc trưng TF-IDF/BoW), đồng thời có thể điều chỉnh bằng tham số C và lựa chọn kernel phù hợp tối ưu hiệu năng.

Bằng cách thử nghiệm và so sánh các tổ hợp phương pháp (TF-IDF + SVM, TF-IDF + Logistic), đề tài hướng đến xây dựng mô hình phân loại cảm xúc hiệu quả và có tính ứng dụng cao trong các bài toán như phân tích phản hồi khách hàng, đánh giá sản phẩm, và hỗ trợ doanh nghiệp theo dói xu hương hài lòng của người dùng trên sàn thương mại điện tử.

KẾT QUẢ DỰ KIẾN

- **Xây dựng tập dữ liệu cảm xúc tiếng Việt**: Thu thập và tiền xử lý một bộ dữ liệu văn bản tiếng Việt có gán nhãn cảm xúc, phục vụ cho quá trình huấn luyện và đánh giá mô hình.
- **Phát triển và triển khai các mô hình học máy**: Áp dụng các phương pháp như SVM, Logistic Regression kết hợp với các kỹ thuật vector hóa từ TF-IDF để phân loại cảm xúc.
- **Đánh giá hiệu suất mô hình**: So sánh độ chính xác và hiệu quả của các mô hình đã triển khai, từ đó đề xuất mô hình tối ưu nhất cho bài toán nhận diện cảm xúc trong văn bản tiếng Việt.

LỜI CAM ĐOAN

Em xin trân trọng cam kết toàn bộ nội dung, kết quả nghiên cứu trình bày trong đồ án tốt nghiệp này là công trình do chính bản thân tôi thực hiện dưới sự hướng dẫn của giảng viên hướng dẫn. Đồ án được hoàn thành bằng kiến thức, nỗ lực và sự sáng tạo của cá nhân em trong suốt quá trình nghiên cứu, không sao chép từ bất cứ công trình nào khác. Mọi số liệu, kết quả thí nghiệm, phân tích trong đồ án đều được thực hiện một cách trung thực, khách quan và khoa học. Các thông tin, dữ liệu sử dụng đều có nguồn gốc rõ ràng và được xử lý theo phương pháp luận chính xác. Tất cả các nội dung tham khảo từ các nguồn tài liệu, công trình nghiên cứu đã công bố đều được trích dẫn đầy đủ theo đúng quy định về trích dẫn khoa học. Mọi tài liệu tham khảo đều được ghi nhận trong danh mục tài liệu tham khảo với đầy đủ thông tin về tác giả, nguồn, năm lưu hành. Việc sử dụng các ý tưởng, kết quả nghiên cứu của người khác (nếu có) đều được dẫn nguồn minh bạch và chỉ sử dụng ở mức độ tham khảo, có đóng góp phát triển thêm. Em hoàn toàn chịu trách nhiệm về tính chính xác và độ tin cậy của các kết quả trình bày trong đồ án. Nếu phát hiện có bất kỳ sự gian lận, sao chép hoặc vi phạm quy tắc đạo đức học thuật, em sẵn sàng chấp nhận mọi hình thức kỷ luật của nhà trường.

Tác giả ĐATN

Chữ ký

Nguyễn Đức Anh

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành đến các thầy cô giáo đã giảng dạy và tạo điều kiện thuận lợi trong quá trình em học tập tại trường.Em xin bày tỏ sự trân trọng, lòng biết ơn và lời cảm ơn sâu sắc nhất đến thầy TS. Nguyễn Mạnh Hiển, giảng viên bộ môn Mạng và An toàn thông tin, Khoa công nghệ thông tin, trường Đại học Thủy Lợi. Sự tận tình chỉ bảo, hướng dẫn và nhiệt huyết của thầy đã tạo nguồn động lực to lớn cho em, thúc đẩy em không ngừng cố gắng để hoàn thành Đồ án tốt nghiệp một cách chỉnh chu và hoàn thiện nhất. Xin cảm ơn những người đã luôn đồng hành cùng với em trên con đường vừa qua. Cảm ơn gia đình, bạn bè, các anh chị đã luôn giúp đỡ, động viên và ủng hộ em trong quá trình học tập. Đồ án tốt nghiệp được em cố gắng thực hiện và hoàn thành với tất cả sự nỗ lực của bản thân nhưng không thể tránh khỏi được một số thiếu sót, em rất mong sẽ nhận được sự góp ý từ thầy cô, bạn bè và những người quan tâm tới đề tài này để đồ án được hoàn thiện hơn nữa.

Em xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

ĐATN Đồ án tốt nghiệp

KLTN Khóa luận tốt nghiệp

LVTN Luận văn tốt nghiệp

SAV Sentiment Analysis Vietnamese

Word2Vector vector representations of words

CBOW Continuous Bag-of-Words

NLP Natural Language Processing

DANH MỤC HÌNH ẢNH

MỞ ĐẦU

Trong bối cảnh các lĩnh vực công nghệ ngày càng phát triển mạnh mẽ, việc tự động nhận diện cảm xúc trong văn bản tiếng Việt đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực như: quản trị doanh nghiệp, xây dựng và phát triển thương hiệu, chăm sóc khách hàng, khảo sát ý kiến và phân tích đánh giá phản hồi từ người dùng.

Ý kiến và cảm nhận của khách hàng hiện nay đóng vai trò vô cùng quan trọng trong việc định hướng sản phẩm cũng như chiến lược kinh doanh. Do đó, các doanh nghiệp ngày càng có nhu cầu cao trong việc triển khai các hệ thống có khả năng phân tích phản hồi khách hàng một cách tự động và hiệu quả. Thông qua đó, doanh nghiệp có thể nắm bắt được thị hiếu, xu hướng và kỳ vọng của người tiêu dùng, từ đó kịp thời điều chỉnh sản phẩm, nâng cao năng lực cạnh tranh, cũng như thích ứng tốt hơn với những thay đổi liên tục của thị trường.

Từ góc độ nghiên cứu, việc xây dựng một hệ thống có khả năng phân tích cảm xúc trong văn bản tiếng Việt là một trong những hướng đi quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Điều này không chỉ góp phần giải quyết các bài toán thực tiễn mà còn thúc đẩy tiến bộ kỹ thuật trong việc xử lý và hiểu ngữ nghĩa tiếng Việt.

Trong đề tài này, chúng tôi tập trung vào việc xây dựng mô hình phân loại cảm xúc người dùng dựa trên văn bản đánh giá và phản hồi. Cụ thể, các cảm xúc sẽ được chia thành ba nhóm chính và bài toán được triển khai dưới dạng một bài toán phân lớp. Mỗi phản hồi được biểu diễn dưới dạng vector đặc trưng để đưa vào mô hình học máy, nhằm huấn luyện hệ thống có khả năng nhận diện và phân loại cảm xúc của người dùng một cách chính xác.

Tổng quan về đề tài

Trong thời gian gần đây, bài toán phân tích và nhận diện cảm xúc đã trở nên phổ biến hơn khi được ứng dụng rộng rãi vào việc xử lý khối lượng lớn dữ liệu trên các nền tảng truyền thông xã hội như mạng xã hội, diễn đàn trực tuyến, blog cá nhân, wiki, hay các hệ thống cộng tác trực tuyến. Đây là một lĩnh vực quan trọng trong điện toán cảm xúc, với mục tiêu chính là phân loại văn bản (và mở rộng sang các dạng dữ liệu khác như âm thanh, hình ảnh) theo chiều hướng cảm xúc — thông thường là tích cực hoặc tiêu cực hay trung lập.

Bài toán này có mối liên hệ chặt chẽ với các lĩnh vực như truy xuất và tổng hợp thông tin, vì cần thực hiện nhiều bước: thu thập dữ liệu, xử lý và gán nhãn trước khi đưa vào mô hình phân tích. Dù phần lớn các nghiên cứu hiện nay tập trung vào ngôn ngữ tiếng Anh, số lượng công trình nghiên cứu mở rộng sang các ngôn ngữ khác, bao gồm cả tiếng Việt, đang ngày càng gia tăng.

Hệ thống nhận diện cảm xúc hiện có thể chia thành hai nhóm chính dựa trên cách tiếp cận: dựa trên tri thức (knowledge-based) và dựa trên thống kê (statistical-based). Trong nhiều nghiên cứu, bài toán này thường được đơn giản hóa thành bài toán phân lớp. Tuy nhiên, trên thực tế, đây là một nhiệm vụ phức tạp thuộc lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), đòi hỏi phải kết hợp nhiều kỹ thuật hỗ trợ như: nhận diện thực thể có tên, trích xuất khái niệm, phát hiện sự châm biếm, phân tích các khía cạnh, và đặc biệt là phát hiện tính chủ quan trong ngữ cảnh. Việc xác định tính chủ quan là một bước không thể thiếu, bởi vì đa số mô hình hiện nay được thiết kế để phân biệt giữa những phát biểu mang tính cảm xúc và các phát biểu mang tính trung lập.

Một số phương pháp tiêu biểu được áp dụng để xử lý bài toán phân tích cảm xúc bao gồm:

• Phương pháp thủ công (Keyword-based):

Đây là cách tiếp cận truyền thống, dựa vào việc xác định các từ mang sắc thái cảm xúc trong văn bản. Từng từ được gán điểm số tích cực hoặc tiêu cực hoặc trung lập và sau đó được tổng hợp lại để xác định cảm xúc tổng thể. Mặc dù dễ triển khai và tốc độ xử lý nhanh, phương pháp này lại không quan tâm đến ngữ cảnh hoặc trật tự từ, dẫn đến độ chính xác thấp nếu bộ từ điển không đủ mạnh. Ngoài ra, các từ phủ định hoặc mang nghĩa chuyển (ẩn dụ, mỉa mai) có thể khiến kết quả bị sai lệch.

• Phương pháp học máy (Machine Learning):

Một trong những hướng tiếp cận hiện đại là sử dụng các mô hình học máy như Support Vector Machine (SVM), Logistic Regression. Kết hợp với đó là kỹ thuật biểu diễn từ bằng vector (Word Embedding) như Word2Vec, đặc biệt với kiến trúc CBOW (Continuous Bag-of-Words) giúp ánh xạ từ ngữ sang không gian vector liên tục nhằm phản ánh quan hệ ngữ nghĩa giữa các từ.

Đối tượng và phạm vi nghiên cứu

- **Đối tượng nghiên cứu**: Các văn bản tiếng Việt thể hiện cảm xúc, đặc biệt là các phản hồi, đánh giá của người dùng về sản phẩm, dịch vụ. Từ các đặc điểm của văn bản, mô hình sẽ học để xác định cảm xúc tương ứng.
- **Phạm vi nghiên cứu**: Dữ liệu sử dụng để huấn luyện và đánh giá mô hình được thu thập từ các bình luận của trang web https://www.kaggle.com/datasets/linhlpv/vietnamese-sentiment-analyst?select=data+-+data.csv.

Phương pháp nghiên cứu

Đồ án kết hợp giữa nghiên cứu lý thuyết và xây dựng mô hình thực nghiệm:

- **Tìm hiểu và tổng hợp tài liệu** liên quan đến chủ đề như: các nghiên cứu trước đây, mô hình phân tích cảm xúc, các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên.

- **Về mặt lý thuyết**: Nghiên cứu tổng quan về lĩnh vực phân tích cảm xúc trong văn bản tiếng Việt, các phương pháp phổ biến trong nhận diện cảm xúc và một số mô hình tiên tiến được ứng dụng trong các công trình khoa học.
- **Về thực nghiệm**: Tiến hành huấn luyện và đánh giá các mô hình học máy trên tập dữ liệu thực tế; sử dụng công cụ lập trình để xử lý, phân tích, và đánh giá độ chính xác dựa trên ba loại cảm xúc chính là **tích cực**, **tiêu cực** và **trung lập**. Qua đó rút ra nhận xét và đánh giá hiệu quả của từng mô hình cũng như các tham số kỹ thuật đi kèm.

CHƯƠNG 1 TỔNG QUAN VỀ ĐỀ TÀI

1.1 Bối cảnh đề tài

Thương mại điện tử đang trở thành môi trường giao dịch quen thuộc của người tiêu dùng, kéo theo sự gia tăng nhanh chóng của các nội dung phản hồi như đánh giá sao, bình luận, nhận xét về chất lượng sản phẩm và dịch vụ. Với nhà bán hàng và sàn thương mại điện tử, những bình luận này không chỉ là "ý kiến", mà còn là tín hiệu trực tiếp về mức độ hài lòng, nguyên nhân phát sinh khiếu nại, cũng như xu hướng kỳ vọng của khách hàng theo thời gian.

Tuy nhiên, dữ liệu bình luận trên các sàn thường có đặc điểm "nhiễu" và khó xử lý theo cách thủ công. Một sản phẩm có thể nhận hàng nghìn phản hồi trong thời gian ngắn; nếu đọc và tổng hợp bằng tay sẽ tốn nhiều nguồn lực, khó cập nhật kịp và dễ bỏ sót những phản hồi quan trọng. Thêm vào đó, bình luận của người dùng không tuân theo chuẩn viết: có thể thiếu dấu, sai chính tả, dùng từ lóng/teencode, lặp ký tự để nhấn mạnh cảm xúc, hoặc kèm emoji. Nhiều trường hợp cảm xúc còn được thể hiện gián tiếp qua ngữ cảnh (ví dụ khen nhưng mỉa mai, hoặc chê nhưng dùng từ "nhẹ").

Từ nhu cầu thực tế đó, phân tích cảm xúc bình luận (sentiment analysis) trở thành một hướng tiếp cận có giá trị: giúp chuyển dữ liệu chữ thành tín hiệu có thể đo lường như tích cực, tiêu cực hoặc trung tính, từ đó hỗ trợ theo dõi chất lượng sản phẩm/dịch vụ, ưu tiên xử lý các phản hồi tiêu cực, và đánh giá xu hướng hài lòng của khách hàng.

Để thực hiện phân tích cảm xúc ở quy mô lớn, việc ứng dụng Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) là phù hợp vì NLP cho phép máy tính xử lý và khai thác ý nghĩa từ văn bản. Khi kết hợp NLP với các mô hình học máy/học sâu, hệ thống có thể tự động học đặc trưng ngôn ngữ trong bình luận tiếng Việt và đưa ra dự đoán cảm xúc với tốc độ nhanh, đồng thời có khả năng mở rộng theo khối lượng dữ liệu thực tế. Vì vậy, đề tài phân tích cảm xúc bình luận người dùng trên các sàn thương mại điện tử được lựa chọn nhằm đáp ứng nhu cầu tổng hợp phản hồi khách hàng một cách hệ thống, nhanh và có cơ sở định lượng.

1.2 Mục đích đề tài

1.Mục tiêu nghiên cứu

Đề tài hướng đến việc xây dựng quy trình và mô hình có khả năng nhận diện cảm xúc từ bình luận tiếng Việt trên các sàn thương mại điện tử. Các mục tiêu chính gồm:

- Xây dựng quy trình xử lý dữ liệu bình luận: thu thập, làm sạch, chuẩn hóa và chuẩn bị dữ liệu cho huấn luyện.
- Thiết kế phương pháp biểu diễn văn bản và lựa chọn mô hình phân loại phù hợp với dữ liệu tiếng Việt.
- Huấn luyện mô hình phân loại cảm xúc theo nhãn (ví dụ: **tích cực – tiêu cực – trung tính**), sau đó đánh giá bằng các chỉ số phổ biến.
- Phân tích kết quả và đề xuất hướng cải thiện mô hình dựa trên các trường hợp dự đoán sai (phủ định, mỉa mai, từ lóng…).

2. Đối tượng nghiên cứu

Đối tượng nghiên cứu là **các bình luận/đánh giá bằng tiếng Việt của người dùng trên sàn thương mại điện tử**, bao gồm phản hồi về: chất lượng sản phẩm, đóng gói, giao hàng, chăm sóc khách hàng, và trải nghiệm mua sắm.

3. Phạm vi nghiên cứu

- **Ngôn ngữ:** tiếng Việt trong môi trường bình luận trực tuyến.
- **Dữ liệu:** bình luận thu thập từ một hoặc nhiều sàn TMĐT (tùy điều kiện dữ liệu của đề tài).
- **Nhiệm vụ phân loại:** phân loại cảm xúc ở mức bình luận (không bắt buộc đi sâu theo từng khía cạnh như “giá”, “ship”, “chất lượng” nếu không có dữ liệu gán nhãn theo khía cạnh).
- **Giới hạn:** đề tài tập trung vào bài toán phân loại cảm xúc và đánh giá mô hình; các phần như phát hiện mỉa mai nâng cao hoặc phân tích cảm xúc theo khía cạnh có thể được xem là hướng phát triển.

CHƯƠNG 2 CƠ SỞ LÝ THUYẾT

Bài toán nhận diện cảm xúc thuộc nhóm các bài toán phân tích ngữ nghĩa, nơi mà mục tiêu chính là giải mã ý nghĩa ẩn sau từng câu hoặc đoạn văn để xác định sắc thái cảm xúc mà người viết muốn truyền tải. Về bản chất, đây là một dạng bài toán phân loại trong lĩnh vực khai phá dữ liệu, khi mà mỗi đoạn văn bản cần được gán vào một nhóm cảm xúc cụ thể.

Trong nghiên cứu này,tôi đề xuất xây dựng một mô hình có khả năng tự động xác định cảm xúc từ văn bản đầu vào – có thể là một câu đơn, một đoạn hội thoại, hoặc một bài viết hoàn chỉnh. Đầu ra của hệ thống sẽ là nhãn cảm xúc tương ứng, chẳng hạn như tích cực, tiêu cực hoặc trung lập, tùy thuộc vào mức độ phân loại mong muốn.

Đối với những ứng dụng phổ biến như đánh giá sản phẩm, bài toán thường được đơn giản hóa bằng cách chia cảm xúc thành ba nhóm: **tích cực**, **tiêu cực** và **trung lập**, giúp mô hình dễ dàng học và phản hồi với hiệu quả cao hơn. Việc xác định cảm xúc này là nền tảng quan trọng trong nhiều hệ thống tương tác người – máy, đặc biệt là các nền tảng mạng xã hội, chatbot chăm sóc khách hàng, và công cụ phân tích đánh giá sản phẩm.

2.1 Các mô hình học máy sử dụng trong phân loại văn bản

2.1.1 Khái niệm tổng quát

Học máy (Machine Learning) là cho phép máy tính học ra quy luật từ một hay nhiều tập dữ liệu giúp dự đoán hoặc quyết định mà không cần lập trình thủ công. Trong bài toán xử lý ngôn ngữ tự nhiên(NLP) như phân tích cảm xúc thì sẽ chuyển đổi văn bản thành dạng thông số qua các phương pháp biểu diễn đặc trưng như Bag-of-Words hoặc TF-IDF. Sau đó các mô hình sẽ học hàm phân loại f(x) dựa trên các vector đặc trưng nhằm dự đoán cảm xúc của bình luận.

Đối với dữ liệu văn bản, đặc trưng đầu vào thường có số chiều rất lớn và mang tính thưa (sparse), đặc biệt khi sử dụng Bag-of-Words hoặc TF-IDF với n-gram. Trong bối cảnh đó, các mô hình tuyến tính như Logistic Regression và SVM tuyến tính thường cho hiệu quả tốt vì chúng tối ưu trực tiếp trên không gian đặc trưng thưa và ít đòi hỏi tài nguyên tính toán. Ngoài ra, các tham số điều chuẩn (regularization) như C(đối với SVM/Logistic) cho phép kiểm soát mức độ phức tạp của mô hình, giúp hạn chế hiện tượng học quá khớp (overfitting) khi dữ liệu có nhiễu hoặc mất cân bằng nhãn.

Trong đồ án này, việc lựa chọn nhóm mô hình truyền thống (TF-IDF kết hợp SVM/Logistic Regression) còn xuất phát từ tính phù hợp với dữ liệu bình luận thương mại điện tử tiếng Việt: văn bản thường ngắn, giàu từ khóa cảm xúc và có nhiều biến thể không chuẩn (teencode, thiếu dấu, sai chính tả, phủ định). Do đó, hiệu quả của mô hình phụ thuộc đáng kể vào bước tiền xử lý và cách biểu diễn đặc trưng. Việc kết hợp TF-IDF với n-gram giúp mô hình nắm bắt các cụm từ mang nghĩa cảm xúc rõ rệt như "không tốt", "rất hài lòng", "giao hàng chậm", từ đó cải thiện khả năng phân loại so với việc chỉ xét từng từ đơn lẻ.

2.1.2 Một số mô hình học máy phổ biến

Logistic Regression (Hồi quy logistic)

Logistic Regression là mô hình phân loại tuyến tính có giám sát, thường được sử dụng cho bài toán phân loại nhị phân và có thể mở rộng cho đa lớp thông qua chiến lược One-vs-Rest hoặc Softmax (Multinomial Logistic Regression). Ý tưởng chính của Logistic Regression là học một hàm tuyến tính trên vector đặc trưng đầu vào, sau đó đưa qua hàm sigmoid để biến đổi thành xác suất thuộc lớp.

Với dữ liệu đầu vào là vector đặc trưng x, mô hình dự đoán xác suất văn bản thuộc lớp $y = 1$như sau:

$$p(y = 1 \mid x) = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Trong đó wlà vector trọng số và blà hệ số chệch (bias). Sau khi có xác suất, mô hình gán nhãn theo ngưỡng (thường là 0.5) hoặc theo lớp có xác suất cao nhất (đối với đa lớp).

Ưu điểm:

- Huấn luyện nhanh, ổn định, phù hợp với dữ liệu đặc trưng thưa và số chiều lớn như TF-IDF.
- Có thể diễn giải: trọng số wphản ánh mức độ đóng góp của từng từ/đặc trưng vào quyết định phân loại.
- Dễ tối ưu bằng regularization (L1/L2) để giảm overfitting.

Hạn chế:

- Là mô hình tuyến tính nên khó nắm bắt các quan hệ phi tuyến hoặc ngữ cảnh phức tạp (ví dụ mỉa mai, đảo nghĩa nhiều tầng).
- Hiệu quả phụ thuộc mạnh vào chất lượng tiền xử lý và cách chọn đặc trưng (n-gram, lọc từ, chuẩn hóa).

Trong đồ án, Logistic Regression được sử dụng như một mô hình baseline quan trọng để so sánh với các mô hình khác khi kết hợp với TF-IDF.

Support Vector Machine(SVM)

Support Vector Machine là mô hình phân loại có giám sát, hoạt động dựa trên ý tưởng tìm một siêu phẳng (hyperplane) phân tách dữ liệu sao cho **biên phân tách (margin)** giữa các lớp là lớn nhất. Với dữ liệu văn bản dạng TF-IDF, thường sử dụng **Linear SVM** do dữ liệu có số chiều rất lớn và phân tách tuyến tính thường đã đạt kết quả tốt.

Về trực giác, SVM cố gắng tìm siêu phẳng:

$$w^T x + b = 0$$

sao cho các điểm dữ liệu thuộc hai lớp nằm ở hai phía khác nhau và khoảng cách đến siêu phẳng là tối đa. Các điểm "quan trọng" nằm gần biên được gọi là **support vectors** và có vai trò quyết định vị trí siêu phẳng.

Ưu điểm:

- Hoạt động rất hiệu quả trên dữ liệu thưa và số chiều lớn (đặc trưng TF-IDF).
- Có khả năng tổng quát tốt, thường cho kết quả cao trong phân loại văn bản.
- Ít bị ảnh hưởng bởi số lượng đặc trưng lớn so với một số mô hình khác.

Hạn chế:

- Cần điều chỉnh tham số (đặc biệt là C) để cân bằng giữa tối ưu margin và sai số.
- Bản chất SVM không sinh xác suất trực tiếp (với LinearSVC), nếu cần xác suất phải hiệu chỉnh thêm (calibration).
- Với dữ liệu cực lớn, thời gian huấn luyện có thể tăng nếu dùng kernel phi tuyến.

Trong bài toán sentiment, SVM thường được xem là lựa chọn mạnh mẽ khi kết hợp với TF-IDF và n-gram, đặc biệt trong môi trường văn bản ngắn và nhiều nhiễu.

2.2 Biểu diễn văn bản bằng TF-IDF

2.2.1 Khái quát chung

TF-IDF là một kỹ thuật biểu diễn văn bản thành vector số, trong đó mỗi từ (hoặc cụm từ n-gram) được gán một trọng số phản ánh mức độ quan trọng của từ đó đối với một văn bản cụ thể trong toàn bộ tập dữ liệu.TF-IDF đánh giá "giá trị thông tin" của từ đó trên hai yếu tố: tần suất xuất hiện trong văn bản và mức độ phổ biến của từ trong toàn bộ tập văn bản.

Cụ thể TF-IDF là sự kết hợp của TF (Term Frequency) – số lần một từ xuất hiện trong một văn bản, và IDF(Inverse Document Frequency) – độ hiếm của từ đó trên toàn bộ tập dữ liệu. Với IDF, các từ xuất hiện thường xuyên ở nhiều văn bản không mang thông tin phân biệt sẽ bị giảm trọng số, ngược lại các từ mang tính phân biệt sẽ được tang trọng số giúp cho mô hình tập trung vào các dấu hiệu quan trọng hơn.

Việc biểu diễn, TF-IDF biến mỗi văn bản thành một vector có số chiều bằng kích thước từ vựng, trong đó mỗi chiều tương ứng với một từ/n-gram và giá trị tại chiều đó chính là trọng số TF-IDF của từ trong văn bản. Nhờ vậy, TF-IDF có ưu điểm rõ rệt về tính đơn giản, tốc độ, dễ triển khai và hiệu quả trong các bài toán xử lý ngôn ngữ tự nhiên.

2.2.2 Nguyên lý hoạt động

Quá trình tạo ra các vector từ thông qua TF-IDF bao gồm:

1. **Xây dựng từ điển (Vocabulary)**

    - Thu thập tất cả từ xuất hiện trong tập văn bản đầu vào (corpus), sau khi xử lý sơ bộ như loại bỏ dấu câu, chuyển chữ thường …
    - Tạo từ điển gồm V từ điển (mỗi phần tử là một từ/n-gram). Khi đó mỗi văn bản sẽ được biểu diễn dưới dạng một vector có số chiều kích thước từ điển V.

2. **Tính TF- Term Frequency (tần suất từ trong văn bản)**

    - Với mỗi văn bản d, TF đo mức độ xuất hiện của từ t trong chính văn bản đó. TF được tính theo số lần xuất hiện hoặc theo dạng chuẩn hóa theo độ dài văn bản.

$$TF(t,d)=f(t,d)$$

$f(t, d)$: số lần từ txuất hiện trong văn bản d.

3. **Tính IDF- Inverse Document Frequency (độ hiếm của từ trong tập dữ liệu)**

- Với mỗi văn bản, IDF đo mức độ phổ biến của từ t trên toàn bộ tập văn bản D. Những từ xuất hiện ở nhiều văn bản (ví dụ: "là", "và", "có") thường ít giá trị phân biệt nên sẽ có IDF thấp, trong khi các từ hiếm nhưng đặc trưng (ví dụ: "hư", "lỗi", "tệ") sẽ có IDF cao. Công thức thường dùng là:

$$\mathrm{IDF}(t, D) = \log\left(\frac{N}{\mathrm{df}(t)}\right)$$

Trong đó:

Nlà tổng số văn bản trong tập D

df(t)là số văn bản có chứa từ t

4. Tính trọng số TF-IDF và tạo vector đặc trưng

- Trọng số TF-IDF của từ t trong văn bản d được tính bằng tích của TF và IDF:

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

- Mỗi văn bản sẽ được chuyển thành một vector TF-IDF, trong đó mỗi chiều tương ứng một từ/n-gram trong từ điển và giá trị tại chiều đó có trọng số TF-IDF.

Kết quả sau quá trình là một tập các vector TF-IDF (document vectors), có thể sử dụng trực tiếp làm đầu vào cho các mô hình học máy như SVM, Logistic Regression trong các bài toán phân loại văn bản, phân tích cảm xúc,…

2.2.3 Các biến thể của TF-IDF:

Những biến thể phổ biến của TF-IDF bao gồm: **Unigram và N-gram**

TF-IDF có thể biểu diễn theo:

- **Unigram**: từ đơn (1-gram).

- **N-gram**: cụm từ liên tiếp 2–4 từ (bi-gram, tri-gram, …) nhằm giữ ngữ cảnh.

2.2.4 Giới hạn kích thước từ điển (max_features)

Khi số lượng từ/n-gram quá lớn, số chiều vector tăng mạnh và gây tốn tài nguyên. Tham số max_features giới hạn số đặc trưng được giữ lại theo mức độ quan trọng.

Lọc từ theo tần suất xuất hiện (min_df, max_df)

- min_df: loại bỏ các từ quá hiếm (nhiễu, ít giá trị học).
- max_df: loại bỏ các từ quá phổ biến (ít phân biệt lớp).
  Trong hệ thống này:
- min_df=2 loại các từ chỉ xuất hiện 1 lần.
- max_df=0.9 loại các từ xuất hiện ở hơn 90% văn bản.

2.3 Các mô hình nhận diện cảm xúc trong văn bản

2.3.1 Tiếp cận phân tích cảm xúc từ xử lý ngôn ngữ tự nhiên

Các phản hồi, đánh giá hoặc bình luận do người dùng để lại trên các nền tảng trực tuyến, như website bán hàng hay mạng xã hội, là một dạng ngôn ngữ tự nhiên được tạo ra một cách tự phát. Để xử lý dữ liệu này trong phân tích cảm xúc, bước đầu tiên là thu thập tập dữ liệu chứa các nội dung nhận xét, bình luận của người dùng sau khi đã trải nghiệm sản phẩm hoặc dịch vụ.

Sau khi thu thập, dữ liệu cần được **làm sạch**: loại bỏ ký tự đặc biệt, các thành phần không mang ý nghĩa, xử lý lỗi chính tả, chuẩn hóa định dạng ngữ pháp, loại bỏ từ dừng… Đây là giai đoạn tiền xử lý giúp chuẩn bị dữ liệu đầu vào có chất lượng tốt cho mô hình.

Tiếp đến, người nghiên cứu tiến hành khảo sát tổng quan tập dữ liệu – bao gồm xác định kích thước tập dữ liệu, đặc tính phân bố cảm xúc, từ vựng thường gặp, ... Từ đó, chọn lựa các đặc trưng phù hợp để đưa vào mô hình phân tích. Việc xác định đúng các chiều dữ liệu (feature selection) là bước then chốt, bởi các đặc trưng càng mang tính đại diện cao thì mô hình càng dễ đưa ra kết quả chính xác.

Cuối cùng, kết quả phân tích sẽ được đánh giá qua các chỉ số như độ chính xác, F1-score, từ đó làm cơ sở cho việc triển khai ứng dụng vào thực tế hoặc điều chỉnh mô hình để cải thiện hiệu quả.

2.3.2 Tiếp cận phân tích cảm xúc bằng phương pháp học máy

Phân tích cảm xúc có thể được hiểu là quá trình đánh giá và nhận diện thái độ, quan điểm, hay cảm xúc của người dùng thông qua ngôn ngữ trong văn bản. Điều này đã trở thành chủ đề trọng tâm trong nhiều nghiên cứu liên quan đến mạng xã hội và dữ liệu người dùng (Liu, 2012; Pang & Lee, 2008).

Trong bối cảnh số lượng dữ liệu văn bản ngày càng lớn, đặc biệt từ các nền tảng số, việc áp dụng **học máy** (machine learning) – một nhánh quan trọng của trí tuệ nhân tạo – để tự động hóa phân tích cảm xúc là hướng tiếp cận phổ biến.

Học máy được chia thành bốn nhóm chính:

- **Học có giám sát** (Supervised Learning): Hệ thống học từ dữ liệu đã được gán nhãn để dự đoán nhãn cho dữ liệu mới. Bao gồm các bài toán như phân loại (classification) và hồi quy (regression).
- **Học không giám sát** (Unsupervised Learning): Dữ liệu đầu vào không có nhãn. Hệ thống tự phát hiện cấu trúc hoặc mẫu ẩn thông qua phân cụm (clustering) hay phát hiện luật (association rule).
- **Học bán giám sát** (Semi-supervised Learning): Kết hợp giữa hai phương pháp trên, một phần dữ liệu được gán nhãn, phần còn lại không có nhãn.
- **Học tăng cường** (Reinforcement Learning): Mô hình học qua quá trình thử - sai, nhận phần thưởng hoặc hình phạt để điều chỉnh hành vi.

Trong nghiên cứu này, phương pháp **học có giám sát** được lựa chọn để xây dựng hệ thống nhận diện cảm xúc từ các phản hồi của người dùng.

2.3.3 Kiến trúc tổng quát của mô hình

Quá trình nghiên cứu được thực hiện theo các bước cơ bản sau:

1. **Thu thập dữ liệu**: Dữ liệu văn bản được lấy từ nguồn thực tế, cụ thể là từ các bình luận, đánh giá trên trang web. Đây là nguồn dữ liệu chứa nhiều phản hồi thực tế, phù hợp với bài toán phân tích cảm xúc.

2. **Tiền xử lý và gán nhãn**: Trong nghiên cứu này, dữ liệu văn bản được cung cấp dưới dạng đã được tiền xử lý sơ bộ, bao gồm các bước làm sạch như: loại bỏ ký tự đặc biệt, xóa các từ không mang giá trị ngữ nghĩa (stop words), chuẩn hóa dấu câu và định dạng văn bản.

    - Để chuẩn bị cho bước vector hóa và huấn luyện mô hình, tôi thực hiện thêm một bước **chuẩn hóa từ vựng**, cụ thể là thay thế khoảng trắng giữa các từ trong cụm bằng dấu gạch dưới (_). Ví dụ: cụm từ "tiếng Việt" được chuyển thành "tiếng_Việt". Việc này giúp mô hình phân tích ngữ nghĩa chính xác hơn, đặc biệt trong các kỹ thuật trích xuất đặc trưng như TF-IDF, nơi các từ ghép cần được giữ nguyên làm một thực thể thống nhất.

    - Sau khi hoàn tất, các câu văn trong tập dữ liệu sẽ có định dạng rõ ràng, nhất quán và thuận lợi cho việc tách từ và huấn luyện mô hình phân loại cảm xúc. Cuối cùng, từng file liệu được gán nhãn cảm xúc tương ứng (ví dụ: "tích cực" hoặc "tiêu cực"), phục vụ cho bài toán học có giám sát.

3. **Chia dữ liệu huấn luyện và kiểm thử**:

    - **Tập huấn luyện**: Được sử dụng để đào tạo mô hình học máy. Các văn bản sau tiền xử lý được biến đổi thành các vector đặc trưng thông qua các phương pháp như TF-IDF, Word2Vec... Mô hình học từ mối quan hệ giữa đặc trưng và nhãn để tìm ra quy luật phân loại.

    - **Tập kiểm thử**: Sau khi mô hình học xong, tập dữ liệu kiểm thử được sử dụng để đánh giá khả năng dự đoán của mô hình với dữ liệu chưa từng thấy. Dữ liệu trong tập này cũng đã được tiền xử lý tiếp đó được chuyển

về vector đặc trưng như tập huấn luyện, sau đó đưa vào mô hình để dự đoán nhãn cảm xúc.

Việc đánh giá hiệu suất mô hình được thực hiện qua các chỉ số đo lường chất lượng như: độ chính xác (accuracy), độ thu hồi (recall), độ chính xác (precision), F1-score...

CHƯƠNG 3 NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT

3.1 Tiền xử lý dữ liệu văn bản

Trong lĩnh vực ngôn ngữ học, "ngôn ngữ tự nhiên" được hiểu là các loại ngôn ngữ hình thành một cách tự nhiên trong quá trình tiến hóa và giao tiếp của loài người, không phải là kết quả của sự lập trình hay thiết kế có chủ đích. Đây là phương tiện giao tiếp cơ bản giữa con người với nhau và có thể tồn tại dưới nhiều hình thức khác nhau như lời nói, chữ viết, ngôn ngữ ký hiệu hay các dạng cảm nhận xúc giác.

Trong bài toán xử lý ngôn ngữ tự nhiên tiếng Việt, trước khi dữ liệu được đưa vào mô hình học máy, một bước quan trọng không thể bỏ qua là **tách từ** và **chuẩn hóa từ ngữ**. Đây là khâu tiền xử lý đóng vai trò nền tảng, giúp mô hình hiểu đúng ngữ nghĩa văn bản, từ đó nâng cao chất lượng của các tác vụ như phân tích cảm xúc, phân loại, tóm tắt hay dịch tự động.

Trong nghiên cứu này, dữ liệu được thu thập từ nguồn đã qua bước làm sạch sơ bộ (xóa HTML, icon, ký tự đặc biệt, dòng trống...). Việc xử lý tập trung chủ yếu vào chuẩn hóa từ ngữ và chính tả, thông qua từ điển ánh xạ các từ viết tắt phổ biến về dạng chuẩn. Việc này được thực hiện bán tự động bằng script Python kết hợp với xử lý chuỗi.

3.1.1 Tách từ

Tách từ (word segmentation) là một bước không thể thiếu trong tiền xử lý ngữ liệu tiếng Việt. Khác với tiếng Anh – nơi dấu cách được dùng để phân biệt ranh giới từ – thì trong tiếng Việt, dấu cách lại chỉ đóng vai trò phân tách các **âm tiết**, chứ không phân định từ hoàn chỉnh. Điều này khiến cho việc tách từ trong tiếng Việt trở nên phức tạp và là một thách thức trong các bài toán xử lý ngôn ngữ tự nhiên.

Ví dụ, cụm từ "đất nước" bao gồm hai âm tiết "đất" và "nước", mỗi âm tiết đều có thể tồn tại độc lập với nghĩa riêng. Tuy nhiên, khi ghép lại, chúng tạo thành một từ mang ý nghĩa khác biệt. Nếu không tách đúng cụm "đất nước", mô hình học máy có thể hiểu nhầm hoặc xử lý sai ngữ cảnh.

Mục tiêu và vai trò

Tách từ giúp hệ thống:

- Hiểu rõ ranh giới từ vựng.
- Loại bỏ nhập nhằng về mặt ngữ nghĩa.
- Tăng độ chính xác khi chuyển văn bản thành các đặc trưng đầu vào cho mô hình học máy.

Một ví dụ minh họa rõ nét cho vấn đề nhập nhằng ngữ nghĩa nếu tách sai:

- "Ăn cơm không được uống rượu." có thể tách thành:
    - Ăn / cơm / không / được / uống / rượu.
    - Ăn / cơm không / được / uống / rượu.

Tùy vào cách tách, nghĩa của câu sẽ thay đổi hoàn toàn.

Quy tắc biểu diễn sau khi tách

Sau khi tách, các từ ghép sẽ được nối bằng ký tự gạch dưới (_) để biểu diễn như một cụm từ thống nhất. Ví dụ: "ẩm thực Việt Nam" → ẩm_thực Việt_Nam

Đây là quy ước phổ biến trong các hệ thống xử lý ngôn ngữ tự nhiên tiếng Việt, nhằm giữ được mạch ngữ nghĩa cho câu.

Các phương pháp tách từ phổ biến

Hiện nay có bốn hướng tiếp cận chính trong việc tách từ tiếng Việt:

1. **Dựa vào từ điển** – So khớp các chuỗi từ trong văn bản với từ điển từ vựng có sẵn.
2. **Dựa vào thống kê** – Sử dụng tần suất xuất hiện của các cụm từ trong tập dữ liệu để xác định ranh giới từ.

3. **Kết hợp từ điển và thống kê** – Phương pháp kết hợp giúp tận dụng cả ưu điểm của từ điển và tính thực tiễn từ dữ liệu.

4. **Dựa trên ký tự (n-gram)** – Tách câu thành các chuỗi ký tự dài cố định như unigram, bigram... Dễ triển khai nhưng hạn chế về mặt ngữ nghĩa.

Trong nghiên cứu này, quá trình tách từ được thực hiện thông qua công cụ word_tokenize của thư viện underthesea. Việc sử dụng công cụ xác định ranh giới từ, tăng độ chính xác và nhất quán cho toàn bộ tập dữ liệu.

3.1.2 Chuẩn hóa từ ngữ và chính tả

Bên cạnh tách từ, bước **chuẩn hóa từ ngữ và chính tả** cũng giữ vai trò quan trọng trong việc làm sạch và làm giàu dữ liệu đầu vào. Trong môi trường mạng xã hội, blog, hoặc các nền tảng bình luận, người dùng thường xuyên sử dụng **viết tắt**, **teencode**, hoặc **sai chính tả** – nếu không xử lý, mô hình học máy sẽ xem đây là các từ độc lập, gây nhiễu và làm giảm độ chính xác khi huấn luyện và dự đoán.

Tại sao cần chuẩn hóa?

Cùng một từ ngữ có thể được viết dưới nhiều dạng khác nhau, chẳng hạn:

- không có thể được viết thành: ko, k, kh
- với → vs
- rồi → r, rùi, r
- cũng → cx
- như thế nào → ntn
- gì → j
- thanks→ cảm ơn
- okie/oke→ok

Nếu giữ nguyên các dạng biểu diễn như vậy, dữ liệu sẽ bị phân mảnh, mô hình khó học được các đặc trưng nhất quán. Việc chuẩn hóa đưa mọi từ viết tắt, sai chính tả, hoặc biểu hiện không chính quy về một **dạng chuẩn thống nhất**, giúp:

- Giảm số lượng từ vựng không cần thiết.
- Tránh hiểu nhầm ngữ nghĩa.
- Tối ưu dung lượng lưu trữ và tốc độ xử lý.

Sau đó văn bản được quét và thay thế từ sai thành dạng chuẩn. Điều này giúp dữ liệu đầu vào nhất quán, giúp mô hình TF-IDF trích xuất đặc trưng ổn định hơn, tăng chất lượng ma trận từ-tần suất và hỗ trợ mô hình hiệu quả hơn.

3.2 Vector hóa văn bản

Trong các hệ thống phân tích cảm xúc sử dụng mô hình học máy, một yêu cầu tiên quyết là dữ liệu văn bản phải được chuyển đổi thành dạng số để máy tính có thể xử lý. Bởi lẽ, các thuật toán không thể làm việc trực tiếp với từ ngữ hay câu chữ như con người, mà chỉ hiểu và tính toán trên các con số.

Quá trình ánh xạ các đơn vị từ vựng (như từ, cụm từ) sang không gian số được gọi là vector hóa văn bản. Đây là bước quan trọng giúp biến đổi văn bản thành các vector đặc trưng trong không gian nhiều chiều, từ đó phản ánh mức độ xuất hiện và vai trò của từ trong toàn bộ tập dữ liệu.

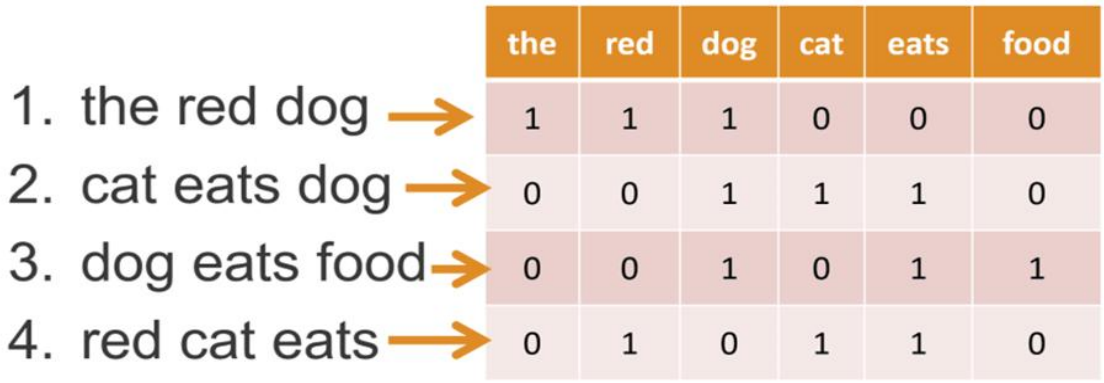
Trong nghiên cứu này, chúng tôi sử dụng phương pháp **TF-IDF (Term Frequency – Inverse Document Frequency)** để vector hóa văn bản. TF-IDF đo lường mức độ quan trọng của một từ dựa trên hai yếu tố: (một) tần suất xuất hiện của từ trong một văn bản cụ thể, và (hai) mức độ "hiếm" của từ đó trong toàn bộ tập văn bản. Nhờ vậy, TF-IDF giúp giảm ảnh hưởng của các từ xuất hiện quá phổ biến và làm nổi bật những từ có khả năng phân biệt cảm xúc tốt hơn, tạo đầu vào phù hợp cho các mô hình học máy như SVM trong bài toán phân loại sentiment.

3.2.1 Các phương pháp biểu diễn văn bản cổ điển (Classic Word Embedding)

• Bag of Words (BoW)

Phương pháp Bag of Words (túi từ) là cách tiếp cận đơn giản nhưng hiệu quả ở mức cơ bản. Mỗi văn bản được biểu diễn bằng một vector có số chiều tương ứng với kích thước của bộ từ vựng. Nếu một từ xuất hiện trong văn bản, vị trí tương ứng trong vector sẽ được gán giá trị 1, nếu không thì gán 0.

Tuy nhiên, mô hình này không tính đến tần suất xuất hiện hoặc ngữ cảnh sử dụng từ, nên thường không phản ánh được ý nghĩa đầy đủ.



Hình 3.1 Mô hình biểu diễn Bag of Words

• TF-IDF (Term Frequency - Inverse Document Frequency)

TF-IDF là phương pháp cải tiến từ BoW, giúp đánh giá **tầm quan trọng** của một từ trong một văn bản so với toàn bộ tập tài liệu. Cách làm này không chỉ quan tâm đến sự xuất hiện, mà còn xét đến sự hiếm gặp của từ đó trong toàn bộ tập dữ liệu – từ càng hiếm nhưng xuất hiện nhiều trong một văn bản thì càng có trọng số cao.

Công thức tính:

- TF (Term frequency): Tần suất xuất hiện của một từ trong một văn bản.

Công thức:

$$tf_i = n_i / N_i$$

Trong đó:

- n_i: số lần từ i xuất hiện trong văn bản
- N_i: tổng số từ trong văn bản

- IDF (Invert Document Frequency): Hay tần số văn bản nghịch đảo, được dùng để đánh giá tầm quan trọng của từ trong văn bản.

  Công thức:

$$idf_i = \log_2 D/d$$

  Trong đó:

  - D: tổng số văn bản có trong tập dữ liệu (ví dụ: một thư mục chứa 500 file văn bản có định dạng .txt, thì D = 500)
  - d: số văn bản có chứa từ i

- TF-IDF (Term frequency - Invert Document Frequency): Sự kết hợp của tần số từ tf và tần số văn bản nghịch đảo idf.

  Công thức:

$$tf - idf_i = tf_i \times idf_i$$

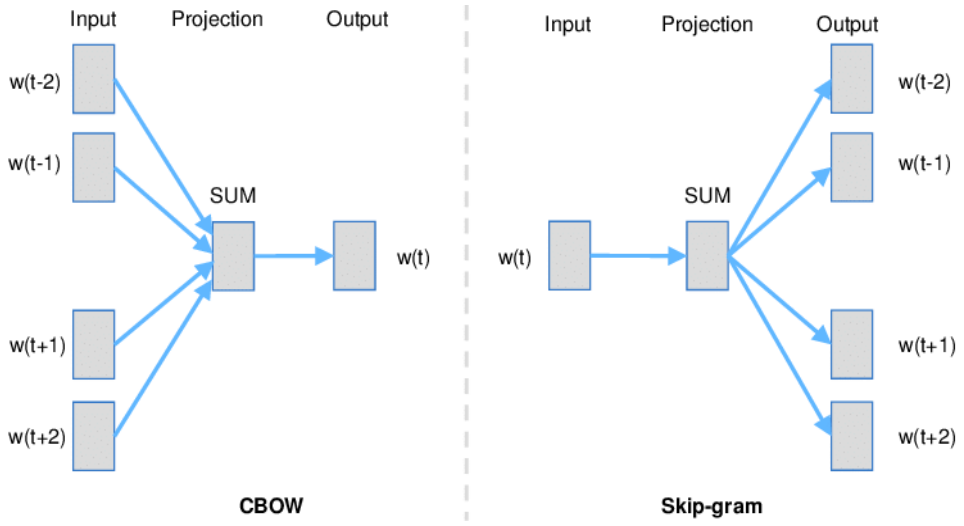
3.2.2 Các phương pháp biểu diễn văn bản hiện đại

• Word2Vec

Word2Vec là một trong những phương pháp embedding hiện đại phổ biến nhất hiện nay. Thay vì xây dựng ma trận đếm hay thống kê thủ công, Word2Vec sử dụng một mạng nơ-ron đơn giản để học biểu diễn vector cho các từ dựa vào ngữ cảnh xuất hiện.

Hai kiến trúc chính:

- **CBOW (Continuous Bag of Words)**: Dùng các từ xung quanh để dự đoán từ trung tâm.
- **Skip-gram**: Dùng một từ để dự đoán các từ xung quanh nó trong một cửa sổ ngữ cảnh.



Hình 3.2 Mô hình CBOW và Skip-Gram

Skip-gram có ưu điểm học được tốt cả với những từ ít xuất hiện nhưng huấn luyện tốn thời gian hơn so với CBOW.

• FastText

FastText là phiên bản cải tiến của Word2Vec, được phát triển bởi Facebook AI. Khác với Word2Vec – học embedding cho toàn bộ từ – thì FastText chia từ thành các đơn vị nhỏ hơn gọi là n-gram ký tự (ví dụ: ng, gra, ram). Việc học biểu diễn dựa trên các n-gram giúp mô hình xử lý tốt hơn các từ hiếm, từ chưa thấy hoặc từ viết sai chính tả.

Điểm mạnh của FastText là giúp tăng khả năng khái quát và cải thiện đáng kể độ chính xác của các mô hình xử lý ngôn ngữ tự nhiên, đặc biệt là tiếng Việt – nơi có nhiều biến thể từ và lỗi chính tả thường gặp.

3.3 Mô hình nhận diện cảm xúc sử dụng học máy

Để giải quyết bài toán phân loại cảm xúc trong văn bản tiếng Việt, đồ án này áp dụng các phương pháp học sâu và học máy cổ điển kết hợp. Cụ thể, chúng tôi sử dụng mô hình học máy Support Vector Machine(SVM) và Logistic Regression nhằm xử lý chuỗi văn bản có tính tuần tự.

Quá trình huấn luyện các mô hình này sử dụng hai phương pháp biểu diễn văn bản chính là **TF-IDF**. Với TF-IDF giúp gán trọng số cho từng từ theo mức độ quan trọng trong toàn bộ tập dữ liệu.

Để mô hình đạt được hiệu quả tốt, dữ liệu đầu vào cần có số lượng đủ lớn và đã được gán nhãn rõ ràng theo từng loại cảm xúc (ví dụ: tích cực, tiêu cực trung lập). Một phần dữ liệu được sử dụng để huấn luyện, phần còn lại dùng để đánh giá độ chính xác của các mô hình.

CHƯƠNG 4 THỰC NGHIỆM

4.1 Xây dựng dữ liệu

Trong lĩnh vực ngôn ngữ học, "ngôn ngữ tự nhiên" được hiểu là các loại ngôn ngữ hình thành một cách tự nhiên trong quá trình tiến hóa và giao tiếp của loài người, không phải là kết quả của sự lập trình hay thiết kế có chủ đích. Đây là phương tiện giao tiếp cơ bản giữa con người với nhau và có thể tồn tại dưới nhiều hình thức khác nhau như lời nói, chữ viết, ngôn ngữ ký hiệu hay các dạng cảm nhận xúc giác.

4.1.1 Bối cảnh và vai trò của dữ liệu huấn luyện

Trong lĩnh vực học máy (machine learning), dữ liệu không chỉ là yếu tố đầu vào – mà còn là nền móng cốt lõi quyết định mức độ thành công của mô hình. Các mô hình học máy thường chứa rất nhiều tham số cần được tối ưu hóa qua nhiều vòng lặp. Do đó, yêu cầu về một tập dữ liệu huấn luyện có kích thước đủ lớn, chất lượng tốt và mang tính đa dạng cao là điều tất yếu.

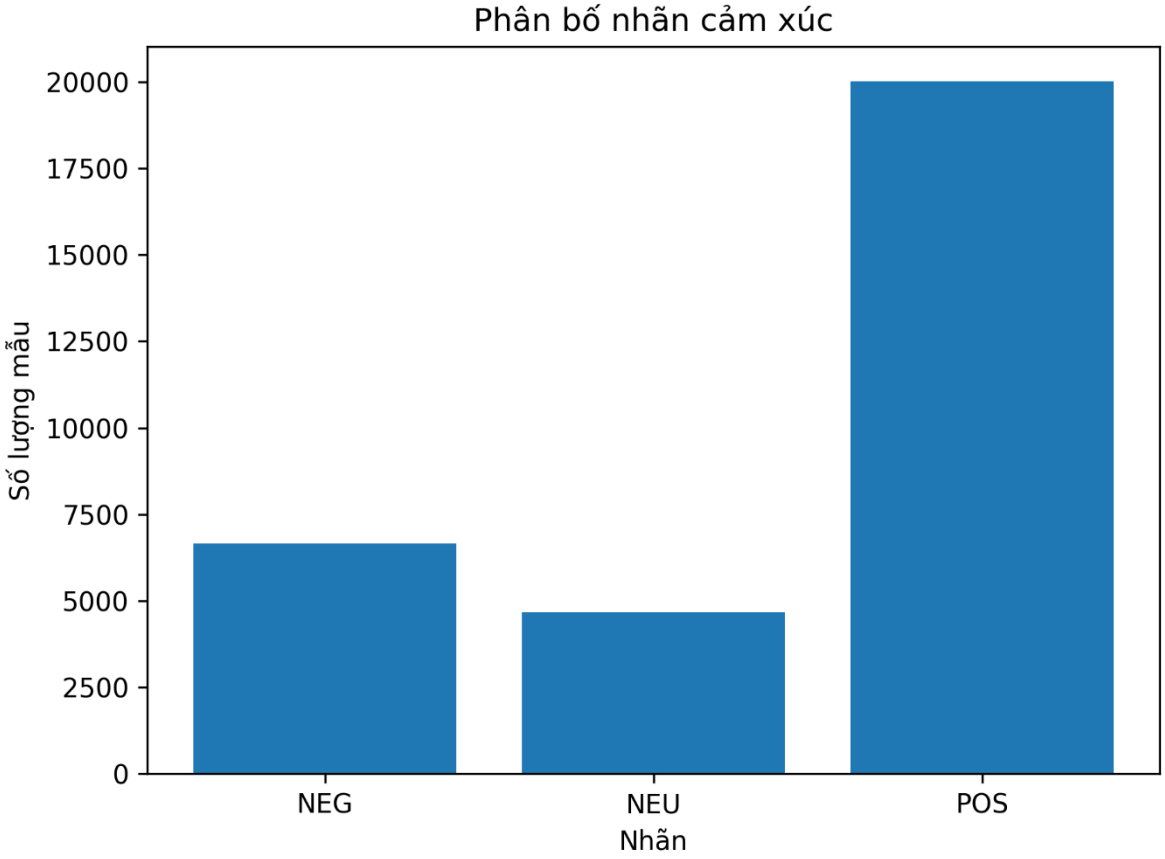
Tuy nhiên, điều này lại đặt ra không ít thách thức cho các cá nhân hoặc nhóm nghiên cứu không trực thuộc các tổ chức có sẵn hệ sinh thái dữ liệu lớn (Big Data). Việc tự thu thập, gán nhãn và xử lý dữ liệu là công việc mất nhiều thời gian, công sức và tài nguyên. Theo kinh nghiệm của chúng tôi, có thể khẳng định rằng hơn 70–80% thời lượng nghiên cứu cho một dự án học máy thường được dành cho khâu xử lý và chuẩn bị dữ liệu. Điều đó càng nhấn mạnh vai trò thiết yếu của việc xây dựng một tập dữ liệu huấn luyện phù hợp và có tính ứng dụng cao trong thực tiễn.

4.1.2 Bộ dữ liệu nhận xét được gán nhãn cảm xúc tích cực, tiêu cực hay trung lập

Dựa trên định hướng ứng dụng vào bài toán phân tích cảm xúc trong ngôn ngữ tiếng Việt, tôi đã chủ động tìm kiếm một bộ dữ liệu chứa các phản hồi, đánh giá thực tế của người tiêu dùng từ các nền tảng thương mại điện tử và bộ dữ liệu này được lấy từ các bình luận của trang web. Đây là một trong những môi trường thực tế chứa lượng lớn bình luận, mang đậm tính chủ quan và cảm xúc của người dùng – một nguồn ngữ liệu lý tưởng để khai thác cho mục tiêu nghiên cứu.

Việc gán nhãn dữ liệu được tiến hành theo cách xem xét dựa trên **mức độ hài lòng thông qua số điểm trung bình của mỗi bình luận**:

- Những phản hồi có điểm đánh giá từ **1 đến 2 điểm** được xếp vào nhóm **tích cực**.
- Những phản hồi có điểm đánh giá **3 điểm** được xếp vào nhóm **trung lập**.
- Những phản hồi có điểm đánh giá từ **4 đến 5 điểm** được xếp vào nhóm **tiêu cực**.



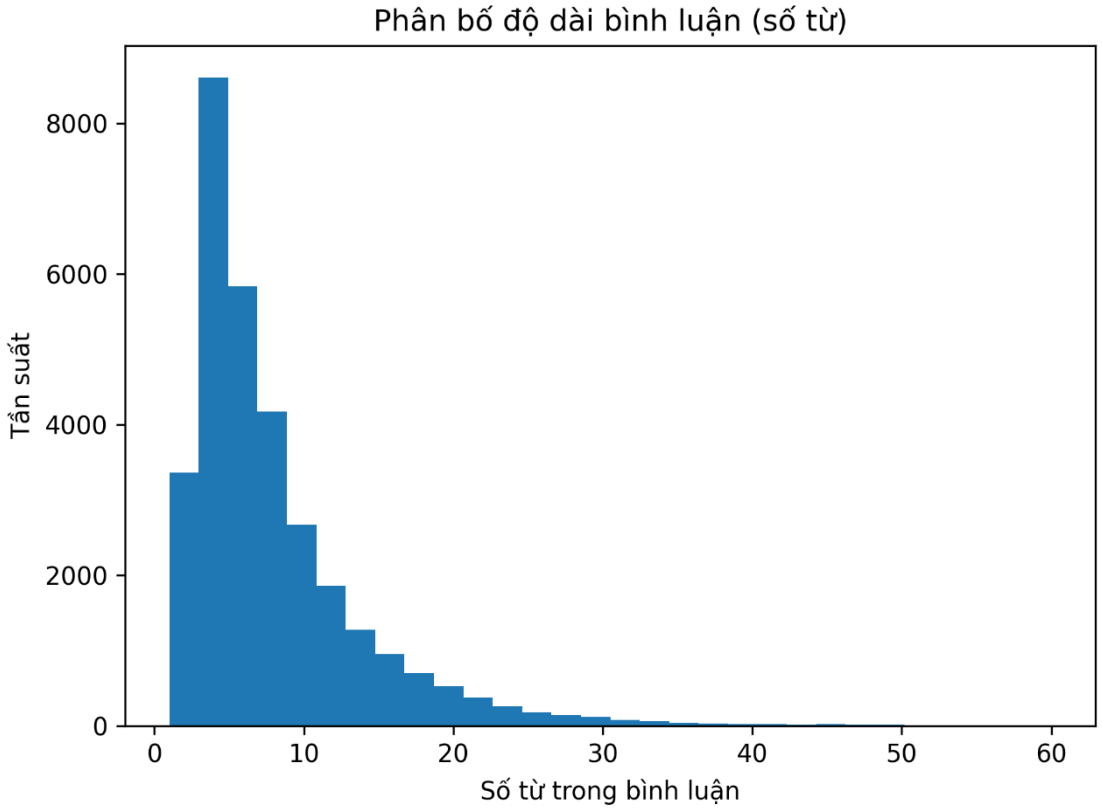
Hình 4.1 Phân bố nhãn của cảm xúc của tập dữ liệu

	comment
2	Áo bao đẹp ạ!!
3	Tuyệt vời !
4	2day ao khong giong trong.
5	Mùi thơm,bôi lên da mềm da.
6	Vải đẹp, dày dặn.
7	Hàng rất đẹp, rất chi là ưng ý.
8	Chất lượng sản phẩm tốt, date dài.
9	Ăn nói và thái độ phục vụ tốt.
10	Đóng gói sản phẩm chắc chắn.
11	tất sờn hết ca chưa dùng mà vay r.
12	Shop phục vụ rất tốt.
13	Mặc thì cũng đc.
14	Chất vải khỏi chê.
15	Thời gian giao hàng rất nhanh.
16	Chất lượng sản phẩm tuyệt vời.
17	vải hơi thô cứng Thời gian giao hàng nhanh.
18	Chất lượng sp chưa thật sự đẹp nhe shop!
19	Rất đáng tiền Thời gian giao hàng rất nhanh.
20	Quần rất đẹp mặc vừa vặn.
21	Cảm giác mua hàng bị hớ thật tệ!
22	Khi mua về nên đi sửa lại.
23	Với giá này thì sản phẩm tạm ổn chưa đc gọi là đẹp lắm.
24	Rất đáng tiền Thời gian giao hàng rất nhanh Chất lượng sản phẩm tuyệt vời.
25	Giá cả chấp nhận được.
26	Nchung là rất ổn 💚.
27	Áo quá đẹp luôn nếu không muốn nói là đẹp, may quá có áo mới đi làm cty mới, shop còn mẫu nào trắng nữa để mình mua.
28	Chưa mặc nhưng thấy chất dày dặn, nhìn chung là hài lòng.

Hình 4.2 Mô tả dữ liệu

Tổng số lượng mẫu trong của cả data_train và data_test là **31.460 mẫu** dữ liệu bình luận:

- **20.093 mẫu tích cực**
- **6669 mẫu tiêu cực**
- **4698 mẫu trung lập**



Hình 4.3 Độ dài của bình luận

4.2 Tiền xử lý dữ liệu

Trong bất kỳ hệ thống học máy nào, đặc biệt là các mô hình xử lý ngôn ngữ tự nhiên (NLP), bước tiền xử lý dữ liệu đầu vào đóng vai trò then chốt trong việc nâng cao hiệu quả và độ chính xác của mô hình. Dữ liệu thô thu thập được thường chứa nhiều thông tin dư thừa, không đồng nhất và có thể gây nhiễu nếu được sử dụng trực tiếp để huấn luyện. Do đó, việc làm sạch và chuẩn hóa dữ liệu là một bước không thể thiếu nhằm đảm bảo chất lượng tập huấn luyện.

Đối với bài toán nhận diện cảm xúc trong văn bản tiếng Việt mà đồ án này hướng tới, dữ liệu đầu vào là các đánh giá, nhận xét thực tế của người dùng về sản phẩm trên các nền tảng thương mại điện tử. Những phản hồi này thường không tuân theo quy chuẩn

văn phong chính thống, đôi khi còn chứa ký hiệu đặc biệt, lỗi chính tả hoặc văn bản không liên quan. Để khắc phục điều đó, quy trình tiền xử lý dữ liệu được xây dựng theo các bước sau:

- **Tách từ**: Đây là bước đặc biệt quan trọng trong xử lý tiếng Việt, do tiếng Việt sử dụng khoảng trắng để phân chia âm tiết chứ không phải ranh giới từ hoàn chỉnh. Việc xác định chính xác ranh giới từ giúp mô hình tránh hiểu sai ngữ nghĩa. Trong nghiên cứu này, chúng tôi sử dụng word_tokenize (thư viện underthesea) để tách từ tự động, đồng thời giữ nguyên quy tắc biểu diễn từ ghép bằng dấu gạch dưới, ví dụ: "ẩm thực Việt Nam" sẽ được biểu diễn thành "ẩm_thực Việt_Nam".
- **Chuẩn hóa văn bản**: Bằng cách ánh xạ các từ viết tắt, teencode hoặc từ sai chính tả về dạng chuẩn, ví dụ: "không" → "ko", "rồi" → "r", "gì" → "j", v.v. Công đoạn này giúp giảm nhiễu và đồng nhất biểu diễn từ trong tập dữ liệu.

Sau khi hoàn tất bước làm sạch văn bản, dữ liệu được chuyển sang giai đoạn **biểu diễn số** để phục vụ huấn luyện mô hình học máy. Trong đồ án này, chúng tôi thực hiện **tách từ (tokenization)** nhằm phân tách câu thành các token (từ hoặc cụm từ) bằng công cụ tách từ tiếng Việt. Trên cơ sở các token đã thu được, văn bản được **vector hóa bằng TF-IDF**, tức là chuyển mỗi văn bản thành một vector đặc trưng phản ánh mức độ quan trọng của từng token trong toàn bộ tập dữ liệu (kết hợp tần suất xuất hiện trong văn bản và mức độ hiếm trong tập văn bản). Nhờ đó, dữ liệu đầu vào trở nên phù hợp để đưa vào mô hình học máy cho bài toán phân loại cảm xúc.

4.3 Huấn luyện mô hình

Trong bất kỳ hệ thống học máy nào, đặc biệt là các mô hình xử lý ngôn ngữ tự nhiên (NLP), bước tiền xử lý dữ liệu đầu vào đóng vai trò then chốt trong việc nâng cao hiệu quả và độ chính xác của mô hình. Dữ liệu thô thu thập được thường chứa nhiều thông tin dư thừa, không đồng nhất và có thể gây nhiễu nếu được sử dụng trực tiếp để huấn luyện. Do đó, việc làm sạch và chuẩn hóa dữ liệu là một bước không thể thiếu nhằm đảm bảo chất lượng tập huấn luyện.

Quá trình xây dựng mô hình nhận diện cảm xúc trong văn bản tiếng Việt được triển khai theo một chuỗi các bước xử lý và huấn luyện mô hình học máy.

Bước 1: Tiền xử lý văn bản thô

Bắt đầu từ tập dữ liệu văn bản tiếng Việt, dữ liệu đầu vào được đưa qua **công cụ tách từ tiếng Việt** (trong đồ án này sử dụng hàm word_tokenize của thư viện **underthesea**). Việc tách từ là bước quan trọng trong chuỗi xử lý ngôn ngữ, giúp xác định ranh giới từ vựng rõ ràng (đặc biệt với tiếng Việt có khoảng trắng theo âm tiết), từ đó tạo tiền đề cho các bước biểu diễn và huấn luyện mô hình phía sau.

Bước 2: Biểu diễn văn bản

Sau khi văn bản được tách từ và chuẩn hóa, hệ thống tiến hành **vector hóa văn bản bằng TF-IDF**. Đây là phương pháp biểu diễn đặc trưng cổ điển nhưng hiệu quả, chuyển mỗi văn bản thành một vector trong không gian nhiều chiều dựa trên **tần suất xuất hiện của từ** trong văn bản và **mức độ hiếm của từ** trong toàn bộ tập dữ liệu. Nhờ cơ chế này, TF-IDF làm nổi bật các từ mang tính phân biệt cao và giảm ảnh hưởng của các từ xuất hiện quá phổ biến.

Bước 3: Huấn luyện mô hình

Từ các đặc trưng TF-IDF đã thu được, hệ thống sử dụng các thuật toán học có giám sát để huấn luyện mô hình phân loại cảm xúc. Trong đồ án này, mô hình trọng tâm là **SVM**, **Logistic Regression** nhận đầu vào là vector TF-IDF và học ranh giới phân tách giữa các lớp cảm xúc (ví dụ: tích cực/tiêu cực hoặc trung lập) dựa trên dữ liệu đã gán nhãn.

Bước 4: Lưu và sử dụng mô hình

Sau khi huấn luyện và đánh giá, mô hình phân loại cảm xúc được **lưu lại** để tái sử dụng cho dữ liệu mới. Mô hình đã lưu có thể được áp dụng để dự đoán cảm xúc của các bình luận/văn bản đầu vào, phục vụ các bài toán như phân tích phản hồi khách hàng, theo dõi chất lượng dịch vụ hoặc hỗ trợ ra quyết định dựa trên dữ liệu văn bản.

Quá trình kiểm tra (test) mô hình nhận diện cảm xúc được thực hiện thông qua một chuỗi xử lý đầu vào văn bản, trong đó dữ liệu test được đưa qua đúng pipeline tiền xử lý và vector hóa giống giai đoạn huấn luyện, sau đó áp dụng mô hình đã lưu để đưa ra dự đoán nhãn cảm xúc.

Bước 1: Văn bản cần được phân tích

Giai đoạn test bắt đầu từ các văn bản tiếng Việt trong tập test (các bình luận/đoạn văn cần phân loại cảm xúc).

Bước 2: Tiền xử lý và tách từ

Tất cả văn bản đầu vào được làm sạch (loại bỏ link, email, số điện thoại, emoji; chuẩn hóa teencode; chuẩn hóa lặp ký tự; lọc ký tự không cần thiết) và sau đó được **tách từ bằng công cụ tách từ tiếng Việt** (trong đồ án này sử dụng word_tokenize của thư viện underthesea). Bước này giúp chuẩn hóa văn bản về dạng token rõ ràng, đảm bảo dữ liệu đầu vào nhất quán trước khi vector hóa.

Bước 3: Biểu diễn văn bản đầu vào bằng TF-IDF

Từ văn bản đã được tách từ, hệ thống sử dụng **mô hình TF-IDF đã huấn luyện từ trước** để chuyển mỗi văn bản thành một vector đặc trưng. Việc dùng lại đúng TF-IDF đã fit ở giai đoạn train giúp đảm bảo cùng không gian đặc trưng và cùng thứ tự các chiều vector giữa train và test.

Bước 4: Dự đoán cảm xúc bằng mô hình đã được huấn luyện (SVM, Logistic Regression)

Vector TF-IDF sau đó được đưa vào **mô hình SVM, Logistic Regression đã huấn luyện** để dự đoán nhãn cảm xúc cho từng văn bản (ví dụ: tích cực/tiêu cực/trung lập). Kết quả dự đoán được xuất ra để phục vụ đánh giá mô hình (accuracy,

precision/recall/F1) hoặc sử dụng trực tiếp trong các bài toán ứng dụng như phân tích phản hồi khách hàng và theo dõi chất lượng dịch vụ.

Bước 5: So sánh và đánh giá kết quả

Để kiểm tra hiệu quả của mô hình, các nhãn dự đoán được **so sánh với nhãn thực tế** trong tập test. Từ đó, hệ thống tính toán và báo cáo các chỉ số đánh giá phổ biến gồm **precision, recall, F1-score** (và có thể kèm theo accuracy). Các chỉ số này được in ra thông qua hàm classification_report của thư viện **scikit-learn**, giúp quan sát chi tiết chất lượng phân loại theo từng lớp cảm xúc.

Ngoài ra, kết quả đánh giá có thể được dùng để **so sánh hiệu năng giữa các mô hình/thiết lập khác nhau** (ví dụ các cấu hình SVM khác nhau, Logistic Regression), từ đó lựa chọn phương án có độ chính xác và độ ổn định tốt nhất để áp dụng vào bài toán thực tế.

4.4 Thực nghiệm và đánh giá kết quả

Sau khi hoàn tất bước tiền xử lý dữ liệu, sẽ tiến hành thực nghiệm mô hình trên hai phương pháp biểu diễn văn bản: **TF-IDF**. Mỗi phương pháp được kết hợp với các thuật toán phân loại khác nhau nhằm so sánh, đánh giá hiệu quả nhận diện cảm xúc trong văn bản tiếng Việt.

Hai thuật toán học máy SVM và Logictic – được sử dụng để huấn luyện mô hình trên tập dữ liệu đã chuẩn hóa. Trong từng lần thực nghiệm, các chỉ số như **độ chính xác phân loại**, **thời gian huấn luyện**, và **tốc độ dự đoán** đều được ghi nhận nhằm phục vụ việc đánh giá khách quan và đề xuất cải tiến cho các giai đoạn nghiên cứu tiếp theo.

Việc huấn luyện và kiểm thử mô hình được triển khai trong môi trường lập trình **Python**, sử dụng các thư viện học máy phổ biến như **Pandas, NumPy và Scikit-learn**. Toàn bộ thí nghiệm được thực hiện trên thiết bị có thông số kỹ thuật như sau:

- **Thiết bị sử dụng**: Lenovo Thinkpad
- **Bộ xử lý**: Intel Core I7-11850H
- **Ổ cứng lưu trữ**: SSD 512 GB
- **Bộ nhớ RAM**: DDR4 32 GB.
- **Ngôn ngữ lập trình**: Python

Với cấu hình như trên, thời gian huấn luyện mô hình có thể không tối ưu như trên các hệ thống GPU hiện đại, nhưng vẫn đảm bảo đủ khả năng đánh giá và kiểm thử hiệu năng của mô hình trên quy mô dữ liệu vừa phải.

Bảng 4.3 Danh sách các mô hình vector hóa và thuật toán phân loại được sử dụng trong thực nghiệm

Thứ tự	**Mô hình vector hóa**	**Phương pháp thực hiện**
1	TF-IDF	SVM
2	TF-IDF	Logistic Regression

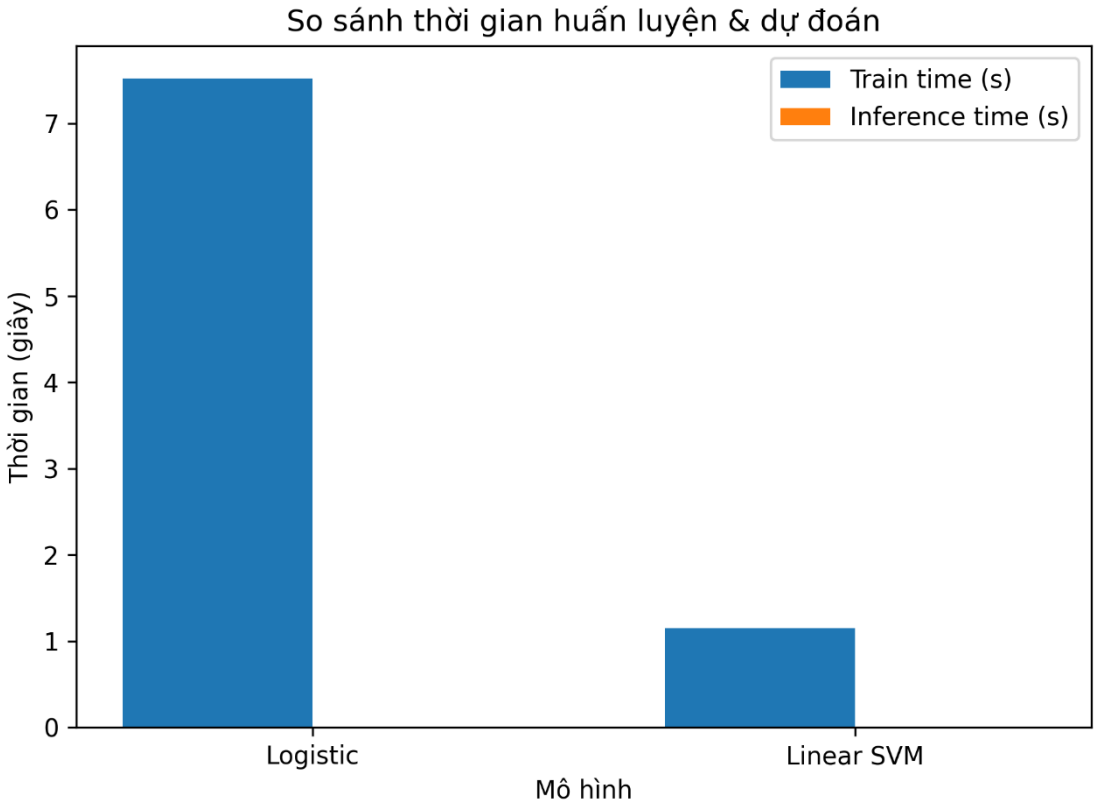
Sau giai đoạn tiền xử lý và vector hóa dữ liệu bằng TF-IDF, chúng tôi xây dựng mô hình học máy các lớp chính như sau:

Huấn luyện mô hình:

- Tách tập train/validation: Tỷ lệ validation 20%

Phân loại bằng mô hình học máy SVM và Logistic Regression (trên tập dữ liệu kiểm tra):

Thời gian chạy:



Hình 4.4 So sánh thời gian chạy của phương pháp

2.Các chỉ số như **precision, recall, F1-score** được tính toán và in ra thông qua hàm classification_report trong thư viện sklearn.

Kết quả thay đổi theo tham số C:

Support Vector Machine(SVM):

C	Accuracy	Macro F1
0.01	0.4427(44.27%)	0.4612
0.1	0.6902(69.02%)	0.6287
1	0.7393(73.93%)	0.6485
10	0.7248(72.48%)	0.6109
100	0.7104(71.04%)	0.5953

Khi **C quá nhỏ (0.01)**, mô hình bị **underfitting**, Accuracy và Macro F1 giảm mạnh.

Khi tăng C lên **0.1 → 1**, chất lượng mô hình cải thiện rõ rệt, trong đó **C = 1** cho kết quả tốt nhất theo cả Accuracy và Macro F1.

Khi **C quá lớn (10, 100)**, hiệu năng giảm dần, cho thấy mô hình có xu hướng **overfitting** hoặc giảm khả năng tổng quát hóa.

=> Qua quá trình thử nghiệm tham số C với mô hình SVM tuyến tính (C ∈ {0.01, 0.1, 1, 10, 100}), kết quả cho thấy mô hình đạt hiệu năng tốt nhất tại **C = 1**, với **Accuracy = 73.93%** và **Macro F1 = 0.6486**

```
=== Evaluation report: SVM model ===
              precision    recall  f1-score   support

         NEG       0.66      0.74      0.70      1331
         NEU       0.33      0.47      0.39       933
         POS       0.93      0.80      0.86      4004

    accuracy                           0.74      6268
   macro avg       0.64      0.67      0.65      6268
weighted avg       0.78      0.74      0.76      6268

Train size: 25,070 | Test size: 6,268
Accuracy: 0.7393 (73.93%)
```

Hình4.5 Kết quả trên tập dữ liệu kiểm tra với PP SVM

Logistic Regression:

C	Accuracy	Macro F1
0.01	0.6522(65.22%)	0.5950
0.1	0.7165(71.65%)	0.6415
1	0.7396(73.96%)	0.6514
10	0.7380(73.80%)	0.6337
100	0.7293(72.93%)	0.6208

Khi **C rất nhỏ (0.01)**, mô hình bị **underfitting** do regularization quá mạnh, dẫn đến hiệu năng thấp (**Accuracy 65.22%, Macro F1 0.5950**).

Khi tăng C lên **0.1**, chất lượng cải thiện rõ rệt (**Accuracy 71.65%, Macro F1 0.6415**), cho thấy mô hình bắt đầu học được các đặc trưng cảm xúc tốt hơn.

C = 1 cho kết quả **tốt nhất** trong các cấu hình thử nghiệm, đạt **Accuracy 73.96%** và **Macro F1 0.6514**. Macro F1 cao nhất cho thấy mô hình cân bằng tốt hơn giữa các lớp (NEG/NEU/POS), phù hợp với dữ liệu lệch lớp.

Khi **C tăng lớn (10, 100)**, hiệu năng **không tăng thêm** mà giảm nhẹ (Macro F1 giảm từ 0.6514 xuống 0.6337 và 0.6208). Điều này cho thấy regularization yếu đi khiến mô hình có thể **giảm khả năng tổng quát hóa** (xu hướng overfitting hoặc tập trung nhiều hơn vào lớp phổ biến).

=> Trong phạm vi thử nghiệm, mô hình Logistic Regression với **C = 1** (các tham số còn lại giữ nguyên) là lựa chọn phù hợp nhất vì đạt **Accuracy cao nhất** và **Macro F1 cao nhất**, thể hiện hiệu năng tổng quát và cân bằng lớp tốt hơn so với các giá trị C còn lại.

```
=== Evaluation report: Logistic Regression model ===
              precision    recall  f1-score   support

         NEG       0.67      0.74      0.71      1331
         NEU       0.33      0.47      0.39       933
         POS       0.93      0.80      0.86      4004

    accuracy                           0.74      6268
   macro avg       0.64      0.67      0.65      6268
weighted avg       0.78      0.74      0.76      6268

Train size: 25,070 | Test size: 6,268
Accuracy: 0.7396 (73.96%)
```

Hình 4.6 Kết quả dữ liệu kiểm tra với PP Logistic Regression

Dưới đây là phần so sánh sơ bộ giữa hai mô hình phân loại cảm xúc: **SVM** và **Logistic Regression**, dựa trên các chỉ số đánh giá cơ bản như **precision, recall, f1-score** và **accuracy**.

Về **precision** – tức độ chính xác của mô hình khi dự đoán một nhãn:

- **SVM** đạt precision **0.90** cho nhãn **0** và **0.88** cho nhãn **1**.
- **Logistic Regression** (3 lớp) đạt precision **0.67** cho **NEG**, **0.33** cho **NEU**, và **0.93** cho **POS**.

Có thể thấy Logistic dự đoán **POS rất "mạnh"** (precision 0.93), nhưng **NEU là lớp yếu nhất** (precision 0.33), kéo giảm chất lượng tổng thể theo trung bình macro.

Xét đến **recall** – tức khả năng thu hồi (tỷ lệ mẫu đúng được nhận diện đúng):

- **SVM** có recall **0.88** (nhãn 0) và **0.90**.
- **Logistic Regression** có recall **0.74** (NEG), **0.47** (NEU), **0.80** (POS).

Điểm đáng chú ý là Logistic vẫn nhận diện **POS khá tốt** (recall 0.80) nhưng **NEU bị bỏ sót nhiều** (recall 0.47).

Về **f1-score** – thước đo cân bằng giữa precision và recall:

- **SVM** đạt f1-score **0.89** cho cả hai nhãn.
- **Logistic Regression** đạt f1-score **0.71** (NEG), **0.39** (NEU), **0.86** (POS).

Điều này cho thấy Logistic phân loại **POS tốt**, **NEG ở mức trung bình**, nhưng **NEU là điểm yếu rõ rệt** (f1-score 0.39).

Về tổng thể, **accuracy**:

- **SVM** đạt **0.89** trên tập kiểm tra (tổng 10.000 mẫu).
- **Logistic Regression** đạt **0.74** trên tập kiểm tra (6.268 mẫu), với **macro avg f1 = 0.65** và **weighted avg f1 = 0.76**.

Kết luận: **SVM đang cho hiệu năng cao và cân bằng hơn** trên bài toán nhị phân 0/1. Trong khi đó, **Logistic Regression** khi mở rộng sang bài toán **3 lớp (NEG/NEU/POS)** cho thấy độ khó tăng lên rõ rệt, đặc biệt là lớp **NEU** (trung lập) khiến các chỉ số macro giảm mạnh.

4.5 Giao diện thử nghiệm:

Để thuận tiện cho quá trình kiểm thử và minh họa khả năng ứng dụng của mô hình phân tích cảm xúc trong thực tế, nhóm xây dựng một giao diện web đơn giản bằng Streamlit. Giao diện này giúp người dùng nhập bình luận tiếng Việt và nhận kết quả dự đoán cảm xúc ngay lập tức:

Hình 4.7 Giao diện thử nghiệm

KẾT LUẬN

Sau quá trình tìm hiểu tài liệu, khảo sát dữ liệu thực tế, xây dựng chương trình và tiến hành thực nghiệm, dưới sự hướng dẫn của thầy **Nguyễn Mạnh Hiển**, em đã hoàn thành đề tài **phân tích cảm xúc bình luận của người dùng trên các sàn thương mại điện tử** theo hướng **học máy**. Đề tài hướng tới việc tận dụng nguồn phản hồi dạng văn bản rất lớn từ người mua hàng để tự động nhận diện thái độ/cảm xúc trong bình luận, từ đó hỗ trợ tổng hợp ý kiến, đánh giá mức độ hài lòng và cung cấp thông tin tham khảo cho nhà bán hàng hoặc hệ thống thương mại điện tử.

Về mặt triển khai, em đã xây dựng được quy trình xử lý dữ liệu bình luận tiếng Việt tương đối đầy đủ, gồm: thu thập và làm sạch dữ liệu, chuẩn hóa văn bản, tách từ/tiền xử lý, biểu diễn văn bản bằng **TF-IDF**, sau đó huấn luyện mô hình phân loại và đánh giá kết quả. Trong phần thực nghiệm, em lựa chọn hai thuật toán phổ biến là **Logistic Regression** và **SVM**, tiến hành huấn luyện trên tập dữ liệu đã chuẩn bị và so sánh hiệu quả dựa trên các chỉ số đánh giá như **Accuracy, Precision, Recall, F1-score**; đồng thời sử dụng **Confusion Matrix** để quan sát trực quan các trường hợp mô hình dự đoán đúng/sai giữa các lớp cảm xúc.

Kết quả thu được cho thấy hướng tiếp cận sử dụng TF-IDF kết hợp với các mô hình phân loại truyền thống có thể đáp ứng tốt bài toán phân tích cảm xúc trên dữ liệu bình luận ngắn trong thương mại điện tử, đồng thời có ưu điểm về tốc độ và khả năng triển khai. Dù vậy, do đặc trưng ngôn ngữ của bình luận trực tuyến tiếng Việt còn nhiều biến thể (viết tắt, teencode, thiếu dấu, lỗi chính tả, phủ định, cách diễn đạt mỉa mai), mô hình vẫn có thể gặp hạn chế ở một số trường hợp khó, đặc biệt với những bình luận mang tính trung tính hoặc có ngữ cảnh phức tạp. Trong giai đoạn tiếp theo, đề tài có thể được nâng cao bằng việc mở rộng và cân bằng dữ liệu gán nhãn, tăng cường các bước chuẩn hóa tiếng Việt, tối ưu tham số mô hình, cũng như thử nghiệm các phương pháp biểu diễn văn bản và mô hình hiện đại hơn nhằm cải thiện độ chính xác và tính ổn định khi áp dụng vào dữ liệu thực tế.

Hướng phát triển đề tài

Trong phạm vi hiện tại, đề tài đã xây dựng được quy trình xử lý dữ liệu bình luận và thử nghiệm các mô hình phân loại cảm xúc dựa trên TF-IDF kết hợp Logistic Regression và SVM. Tuy nhiên, dữ liệu bình luận trên sàn thương mại điện tử có nhiều đặc thù ngôn ngữ và thay đổi liên tục, vì vậy đề tài có thể tiếp tục phát triển theo các hướng sau:

Thứ nhất, mở rộng và nâng cao chất lượng dữ liệu.

Tập dữ liệu có thể được mở rộng theo nhiều ngành hàng (điện tử, thời trang, mỹ phẩm…), nhiều thời điểm (theo tháng/quý) và nhiều nguồn (nhiều sàn khác nhau). Đồng thời, cần tăng chất lượng nhãn bằng cách bổ sung quy trình gán nhãn chéo (một bình luận do nhiều người gán nhãn) và thống kê mức độ đồng thuận. Việc kiểm tra và xử lý mất cân bằng lớp (nếu có) cũng là hướng cần thực hiện để mô hình học ổn định hơn, đặc biệt với lớp trung tính thường dễ bị nhiễu.

Thứ hai, cải tiến tiền xử lý tiếng Việt theo đặc thù bình luận trực tuyến.

Ngoài các bước chuẩn hóa cơ bản, đề tài có thể phát triển thêm các module xử lý teencode, lỗi chính tả phổ biến, từ lóng, emoji và ký tự kéo dài (ví dụ “đẹpppp”, “okkk”). Bên cạnh đó, xử lý phủ định và cấu trúc đảo nghĩa trong tiếng Việt (ví dụ “không tệ”, “chưa tốt lắm”) là yếu tố ảnh hưởng trực tiếp đến kết quả phân loại cảm xúc, nên cần có các quy tắc hoặc kỹ thuật nâng cao để giảm nhầm lẫn.

Thứ ba, thử nghiệm thêm các kỹ thuật biểu diễn văn bản và tối ưu đặc trưng.

TF-IDF là phương pháp mạnh và hiệu quả cho phân loại văn bản ngắn, tuy nhiên đề tài có thể mở rộng bằng cách thử nghiệm các mức n-gram khác nhau, lựa chọn đặc trưng theo thống kê (Chi-square, Mutual Information) hoặc giảm chiều dữ liệu (Truncated SVD) để tăng tốc và giảm nhiễu. Ngoài ra, có thể so sánh TF-IDF với các dạng embedding như fastText/Word2Vec để xem mức cải thiện trong từng nhóm dữ liệu.

TÀI LIỆU THAM KHẢO

[1].Phạm Đình Khánh, “Hồi qui Logistic”, Deep AI Book – Machine Learning lý thuyết tới thực hành.
[2].Võ Trường Duy, “Bài 1: Giới thiệu về Machine Learning”, Machine Learning cơ bản, 26/12/2016.
[3].“Giới thiệu về Support Vector Machine (SVM)”, Viblo.
[4].“Hồi quy Logistic (Logistic Regression)”, ML Glossary (tiếng Việt).
[5]. Võ Trường Duy, “Bài 10: Logistic Regression”, Machine Learning cơ bản, 27/01/2017.
[6]. Võ Trường Duy, “Bài 19: Support Vector Machine (SVM)”, Machine Learning cơ bản, 09/04/2017.
[7]. Võ Trường Duy, “Bài 21: Kernel Support Vector Machine”, Machine Learning cơ bản, 22/04/2017.
[8]. Võ Trường Duy, “Bài 33: Các phương pháp đánh giá một hệ thống phân lớp (Precision, Recall, F1, …)”, Machine Learning cơ bản, 31/08/2017.
[9]. “TF-IDF (term frequency – inverse document frequency)”, Viblo.
[10]. Phạm Đình Khánh, “Feature Engineering: Bag-of-Words và TF-IDF (thực hành phân loại văn bản)”, Deep AI Book – Machine Learning lý thuyết tới thực hành.

KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

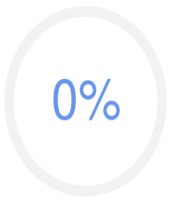
Thông tin tài liệu

Tên tài liệu:	63CNTT4_2151062703_NguyenDucAnh_DATN_v2-content
Tác giả:	Ngành CNTT
Điểm trùng lặp:	9
Thời gian tải lên:	14:22 18/01/2026
Thời gian sinh báo cáo:	14:24 18/01/2026
Các trang kiểm tra:	42/42 trang

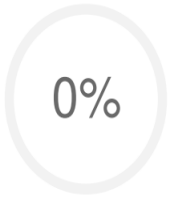
Kết quả kiểm tra trùng lặp

Có 9% nội dung trùng lặp

Có 91% nội dung không trùng lặp



Có 0% nội dung người dùng loại trừ



Có 0% nội dung hệ thống bỏ qua