

TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG

KHOA THƯƠNG MẠI ĐIỆN TỬ



HỌC PHẦN: ĐỀ ÁN THỰC HÀNH 1

**PHÂN TÍCH CẢM XÚC KHÁCH HÀNG TỪ ĐÁNH GIÁ SẢN PHẨM
CỦA CÔNG TY HASAKI BẰNG KỸ THUẬT XỬ LÝ NGÔN
NGỮ TỰ NHIÊN**

Giảng viên hướng dẫn : ThS. Trần Văn Lộc

Lớp: 48K29.2

Nhóm: Nhóm 6

Thành viên : Trương Nguyễn Thành Nhân

Phạm Thị Hồng Thư

Hồ Lê Khôi Nguyên

Đà Nẵng, ngày 08 tháng 05 năm 2025

MỤC LỤC

A. GIỚI THIỆU	5
1. Lý do chọn đề tài	5
2. Mục tiêu cần đạt	6
2.1 Mục tiêu về lý thuyết.	6
2.2 Mục tiêu về ứng dụng	6
3. Mô tả bài toán	7
B. TỔNG QUAN	8
1. Giới thiệu về cửa hàng Hasaki	8
C. CƠ SỞ LÝ THUYẾT	9
1. Khai phá dữ liệu	9
2. Xử lý ngôn ngữ tự nhiên NLP	10
2.1 Quy trình xử lý ngôn ngữ tự nhiên NLP	11
2.2 Ứng dụng NLP trong phân tích cảm xúc khách hàng	12
3. Phân tích cảm xúc	13
4. API Gemini	14
5. Thuật toán	15
6. Nền tảng Streamlit	19
D. PHƯƠNG PHÁP NGHIÊN CỨU	20
1. Cách thức triển khai	20
2. Chi tiết từng module	22
2.1 Module Thu thập và Tiền xử lý dữ liệu	22
2.1.1. Thu thập dữ liệu	22
2.1.2. Tiền xử lý và chuẩn hóa dữ liệu	32
2.2 Module Huấn luyện mô hình	37
2.3 Module Website	45
KẾT LUẬN	48
TÀI LIỆU THAM KHẢO	50

DANH MỤC BẢNG BIỂU

Bảng 1: Thông tin dữ liệu về sản phẩm	29
Bảng 2: Mô tả dữ liệu chứa thông tin về đánh giá của sản phẩm	31

DANH MỤC HÌNH ẢNH

Hình 1: Mô tả bài toán	7
Hình 2: Framework bài toán	20
Hình 3: Giao diện trang thương mại điện tử Hasaki	22
Hình 4: Trang thông tin về sản phẩm.....	23
Hình 5: Trang thông tin chi tiết về sản phẩm.....	24
Hình 6: Phần đánh giá về sản phẩm của khách hàng	24
Hình 7: Repo của dự án.....	25
Hình 8: Dữ liệu thu thập được lưu theo tên có định dạng thời gian	25
Hình 9: Dữ liệu chứa thông tin về sản phẩm thu thập được từ trang Hasaki.....	27
Hình 10: Tổng quan dữ liệu sau khi bóc tách nhãn thực thể	33
Hình 11: Kết quả sau khi gộp comment.....	33
Hình 12: Quy trình tiền xử lý dữ liệu.....	34
Hình 13: Từ điển giải mã một số từ viết tắt	35
Hình 14: Dữ liệu sau khi được tiền xử lý.....	36
Hình 15: Word Cloud.....	36
Hình 16: Kết quả của từng Fold.....	40
Hình 17: Kết quả của từng Fold trực quan.....	41
Hình 18: Tính trung bình độ chính xác và chọn ra mô hình tốt nhất.....	41
Hình 19: Confusion Matrix	42
Hình 20: Classification Report.....	42

Hình 21: ROC và AUC	43
Hình 22: Ví dụ.....	43
Hình 23: Kết quả ví dụ	44
Hình 24: Quy trình xây dựng website	45
Hình 25: Kết quả của Website	47

A. GIỚI THIỆU

1. Lý do chọn đề tài

Trong bối cảnh công nghệ thông tin bùng nổ và mạng xã hội phát triển mạnh mẽ, môi trường kinh doanh hiện nay trở nên cạnh tranh khốc liệt hơn bao giờ hết. Để tồn tại và phát triển, các doanh nghiệp không chỉ cần cung cấp sản phẩm/dịch vụ chất lượng mà còn phải nắm bắt và thấu hiểu nhu cầu, cảm xúc cũng như kỳ vọng ngày càng đa dạng của khách hàng. Những đánh giá của người dùng trên các nền tảng trực tuyến không chỉ phản ánh mức độ hài lòng mà còn là nguồn dữ liệu quý giá thể hiện cảm xúc, kỳ vọng và trải nghiệm thực tế, từ đó ảnh hưởng trực tiếp đến uy tín và hình ảnh thương hiệu.

Đặc biệt trong lĩnh vực làm đẹp và chăm sóc sức khỏe - nơi chất lượng dịch vụ và trải nghiệm khách hàng đóng vai trò then chốt - việc khai thác hiệu quả dữ liệu phản hồi của khách hàng để nâng cao dịch vụ là một yêu cầu cấp thiết. Đà Nẵng, với vai trò là trung tâm du lịch lớn và thị trường tiềm năng trong ngành làm đẹp, chứng kiến sự phát triển mạnh mẽ của các chuỗi cửa hàng mỹ phẩm, spa và dịch vụ chăm sóc da, trong đó Hasaki là một trong những thương hiệu tiêu biểu. Với lưu lượng khách hàng lớn và sự đa dạng về đối tượng phục vụ, chi nhánh Hasaki Đà Nẵng là một trường hợp điển hình để nghiên cứu và phân tích chất lượng dịch vụ dựa trên phản hồi thực tế từ khách hàng.

Việc ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và mô hình học máy như Logistic Regression, Naïve Bayes, Support Vector Machine vào phân tích cảm xúc từ đánh giá của khách hàng sẽ giúp doanh nghiệp hiểu rõ mức độ hài lòng, xác định các điểm mạnh - yếu trong cung cách phục vụ, từ đó đề xuất các giải pháp cải thiện phù hợp. Với những lý do

trên, đề tài **“Phân tích cảm xúc khách hàng từ đánh giá sản phẩm của Công ty Hasaki bằng kỹ thuật xử lý ngôn ngữ tự nhiên”** không chỉ mang tính thời sự cao mà còn có giá trị thực tiễn sâu sắc, đóng góp thiết thực vào việc nâng cao chất lượng dịch vụ trong ngành làm đẹp hiện nay.

2. Mục tiêu cần đạt

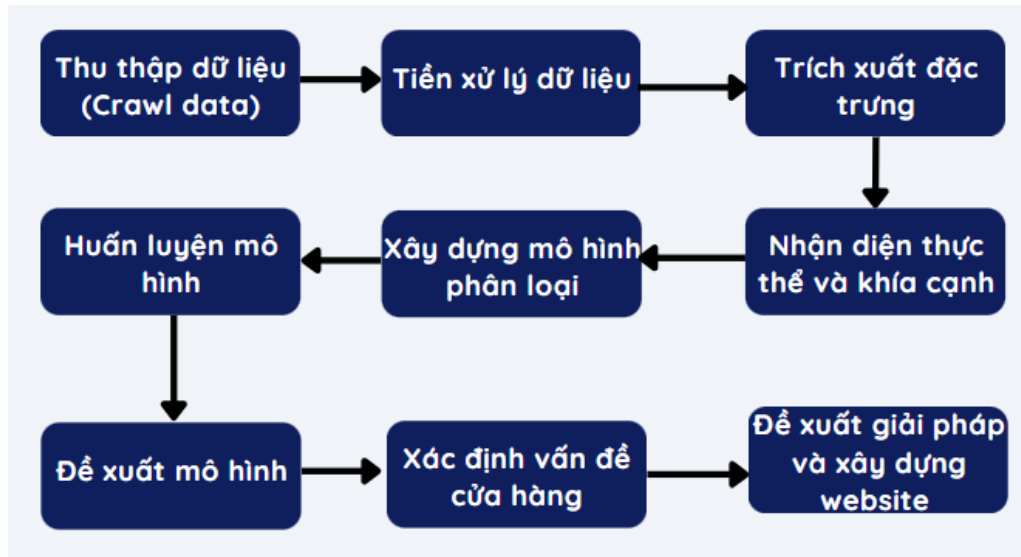
2.1 Mục tiêu về lý thuyết.

- Tìm hiểu phương pháp phân tích cảm xúc dựa đánh giá sản phẩm của người dùng
- Xây dựng mô hình học máy: Xác định các đầu vào cần thiết và xây dựng mô hình học máy

2.2 Mục tiêu về ứng dụng

- Xây dựng mô hình phân tích cảm xúc dựa trên dữ liệu đánh giá sản phẩm của người dùng
- Áp dụng kết quả đánh giá để hỗ trợ cho các nhà quản trị có thể theo dõi tình hình kinh doanh, phản ứng đối với của sản phẩm, và giúp khách hàng ra quyết định mua sắm

3. Mô tả bài toán



Hình 1: Mô tả bài toán

Quy trình phân tích và đánh giá chất lượng dịch vụ của Hasaki gồm những bước chính sau:

- Thu thập đánh giá và bình luận của khách hàng về sản phẩm và dịch vụ tại Hasaki từ các nguồn như website chính thức <https://hasaki.vn/>
- Thực hiện các thao tác tiền xử lý dữ liệu bao gồm tiền xử lý dữ liệu có cấu trúc và xử lý dữ liệu văn bản phi cấu trúc.
- Trích xuất từ khóa.
- Nhận diện thực thể và phân tích cảm xúc bao gồm: chọn nhãn, phân tích nhãn và bóc tách nhãn của cả thực thể và cảm xúc.
- Áp dụng các mô hình học máy như Logistic Regression, Naïve Bayes, hoặc SVM để phân loại cảm xúc của từng bình luận, gán nhãn phù hợp với từng thực thể.

- Xây dựng mô hình phân loại bình luận khách hàng (phân loại cả thực thể và cảm xúc).
- Tiến hành huấn luyện mô hình trên dữ liệu phản hồi thực tế để đạt được độ chính xác cao trong việc phân loại cảm xúc và chọn mô hình tốt nhất.
- Phân tích tập trung vào các điểm yếu hoặc dịch vụ thường xuyên bị phản hồi tiêu cực
- Đưa ra các giải pháp nâng cao chất lượng dịch vụ và xây dựng công cụ trực quan (dashboard/website) hỗ trợ ban quản lý theo dõi, đánh giá hiệu quả cải thiện dịch vụ.

B. TỔNG QUAN

1. Giới thiệu về cửa hàng Hasaki

Hasaki Beauty & Spa là một trong những hệ thống phân phối mỹ phẩm chính hãng, được mỹ phẩm và dịch vụ chăm sóc sắc đẹp hàng đầu tại Việt Nam. Thành lập từ năm 2016, Hasaki không ngừng mở rộng quy mô với hơn 100 chi nhánh trên toàn quốc, cung cấp hơn 10.000 sản phẩm từ các thương hiệu uy tín trong và ngoài nước. Bên cạnh hoạt động bán lẻ mỹ phẩm, Hasaki còn phát triển hệ thống Hasaki Clinic & Spa, chuyên cung cấp các dịch vụ chăm sóc da công nghệ cao theo tiêu chuẩn y khoa.

Hasaki cam kết mang đến cho khách hàng trải nghiệm làm đẹp toàn diện với ba tiêu chí cốt lõi: Chất lượng – Chính hãng – Chi phí hợp lý. Mô hình "Pharma-beauty" kết hợp giữa nhà thuốc – mỹ phẩm – phòng khám da liễu giúp Hasaki tạo được sự khác biệt trên thị trường và chiếm được lòng tin từ hàng triệu khách hàng.

Tại Đà Nẵng, Hasaki đã và đang vận hành các chi nhánh với vị trí thuận lợi, không gian hiện đại và đội ngũ chuyên viên tư vấn tận tâm. Mỗi ngày, Hasaki phục vụ hàng trăm lượt khách hàng với nhu cầu đa dạng: từ mua sắm mỹ phẩm, tư vấn da liễu, đến trị liệu thẩm mỹ chuyên sâu.

Cùng với chiến lược số hóa trải nghiệm người dùng thông qua ứng dụng di động, website thương mại điện tử và hệ thống chăm sóc khách hàng tự động, Hasaki đang từng bước khẳng định vị thế là thương hiệu tiên phong trong ngành làm đẹp tại Việt Nam, đặc biệt tại các thành phố lớn như TP. Hồ Chí Minh, Hà Nội và Đà Nẵng.

C. CƠ SỞ LÝ THUYẾT

1. Khai phá dữ liệu

Khai phá dữ liệu (Data mining)

Khai phá dữ liệu là quá trình sử dụng các kỹ thuật và thuật toán tiên tiến để phân tích và khám phá những thông tin ẩn quý giá từ các tập dữ liệu lớn. Nhờ khai phá dữ liệu, các tổ chức có thể biến đổi kho dữ liệu khổng lồ thành những kiến thức hữu ích, hỗ trợ việc đưa ra quyết định sáng suốt và nâng tầm hiệu quả hoạt động (Data mining là gì?, 2023).

Quy trình khai thác dữ liệu cơ bản gồm 7 bước, bao gồm:

- Làm sạch dữ liệu: Xử lý các lỗi và thiếu sót trong dữ liệu để đảm bảo tính chính xác và nhất quán.
- Tích hợp dữ liệu: Kết hợp dữ liệu từ nhiều nguồn khác nhau thành một tập dữ liệu thống nhất.

- Lựa chọn dữ liệu: Xác định các thuộc tính dữ liệu phù hợp cho việc phân tích.
- Chuyển đổi dữ liệu: Biến đổi dữ liệu sang dạng phù hợp cho các thuật toán khai phá dữ liệu.
- Khai phá dữ liệu: Áp dụng các thuật toán và kỹ thuật để tìm kiếm các mẫu hình, quy luật và mối quan hệ ẩn trong dữ liệu.
- Đánh giá mẫu: Đánh giá độ tin cậy và tính hữu ích của các mẫu hình được phát hiện.
- Trình bày thông tin: Hiển thị kết quả khai phá dữ liệu dưới dạng báo cáo, biểu đồ, hình ảnh dễ hiểu.

Lợi ích của khai phá dữ liệu là khám phá tri thức mới, hỗ trợ ra quyết định, tối ưu hóa hoạt động, tăng lợi thế cạnh tranh, phát triển sản phẩm và dịch vụ mới. Thường được ứng dụng trong bán lẻ, ngân hàng, y tế, viễn thông, chính phủ.

2. Xử lý ngôn ngữ tự nhiên NLP

Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) là một nhánh chuyên biệt của trí tuệ nhân tạo (Artificial Intelligence – AI), tập trung vào việc nghiên cứu và phát triển các hệ thống có khả năng tương tác với ngôn ngữ tự nhiên của con người. NLP cho phép máy tính hiểu, diễn giải, phân tích và tạo ra ngôn ngữ dưới dạng văn bản hoặc lời nói một cách có hệ thống và hiệu quả (Chowdhury, 2003).

Với vai trò là cầu nối giữa giao tiếp ngôn ngữ của con người và khả năng tính toán của máy móc, NLP được ứng dụng rộng rãi trong nhiều lĩnh vực như phân tích dữ liệu văn bản, dịch máy, tóm tắt văn bản tự động,

phân tích cảm xúc, và xây dựng hệ thống hỏi đáp thông minh. Mục tiêu tổng quát của NLP là cung cấp cho máy tính khả năng hiểu và phản hồi ngôn ngữ tự nhiên một cách gần giống như con người (Jurafsky & Martin, 2021).

2.1 Quy trình xử lý ngôn ngữ tự nhiên NLP

Một hệ thống NLP điển hình thường trải qua các giai đoạn sau:

- **Tiền xử lý văn bản (Text Preprocessing):** Là bước đầu tiên nhằm chuẩn hóa dữ liệu văn bản, bao gồm các thao tác như: chuyển văn bản về chữ thường, loại bỏ ký tự đặc biệt, tách từ (tokenization), loại bỏ từ dừng (stop words), và chuẩn hóa từ (stemming, lemmatization). Giai đoạn này đóng vai trò then chốt trong việc đảm bảo chất lượng dữ liệu đầu vào cho mô hình phân tích.
- **Biểu diễn văn bản (Text Representation):** Văn bản sau khi được tiền xử lý cần được chuyển đổi thành dạng số để máy tính có thể xử lý. Các kỹ thuật phổ biến bao gồm Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), và các phương pháp nhúng từ (word embeddings) như Word2Vec, GloVe và BERT.
- **Xây dựng và huấn luyện mô hình:** Dữ liệu sau khi được biểu diễn sẽ được sử dụng để huấn luyện các mô hình học máy hoặc học sâu. Một số mô hình truyền thống thường được sử dụng là Naïve Bayes, Logistic Regression, Support Vector Machine (SVM); trong khi các mô hình hiện đại hơn sử dụng mạng nơ-ron

sâu như RNN, LSTM hoặc Transformer.

- **Đánh giá mô hình:** Mô hình sau khi huấn luyện được đánh giá dựa trên các chỉ số như Accuracy, Precision, Recall và F1-score nhằm đo lường độ chính xác và hiệu quả phân loại.

2.2 Ứng dụng NLP trong phân tích cảm xúc khách hàng

Phân tích cảm xúc (Sentiment Analysis) là một trong những ứng dụng nổi bật của NLP, đặc biệt trong lĩnh vực kinh doanh và thương mại điện tử. Mục tiêu của phân tích cảm xúc là nhận diện và phân loại thái độ của người dùng đối với sản phẩm hoặc dịch vụ thông qua văn bản đánh giá hoặc bình luận.

Việc ứng dụng NLP vào phân tích cảm xúc mang lại nhiều lợi ích thiết thực cho doanh nghiệp, bao gồm:

- Xác định xu hướng hài lòng hoặc không hài lòng của khách hàng;
- Phát hiện kịp thời các vấn đề trong dịch vụ;
- Hỗ trợ ra quyết định dựa trên dữ liệu phản hồi thực tế;
- Tối ưu hóa chiến lược chăm sóc và giữ chân khách hàng.

Theo Liu (2012), các kỹ thuật phân tích cảm xúc có thể được chia thành ba cấp độ: cấp độ tài liệu (document-level), cấp độ câu (sentence-level), và cấp độ khía cạnh (aspect-level). Trong đó, phân tích theo khía cạnh cho phép doanh nghiệp hiểu rõ hơn khách hàng đánh giá tích cực hoặc tiêu cực về yếu tố cụ thể nào như chất lượng sản phẩm, thái độ nhân viên, thời gian giao hàng

3. Phân tích cảm xúc

Phân tích cảm xúc là một nhánh quan trọng của xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), nhằm xác định, diễn giải và phân loại thái độ, quan điểm hoặc cảm xúc của con người được thể hiện qua văn bản. Kỹ thuật này đóng vai trò thiết yếu trong việc giúp doanh nghiệp hiểu rõ hơn về cảm nhận của khách hàng, từ đó đưa ra quyết định phù hợp nhằm nâng cao chất lượng sản phẩm, dịch vụ và trải nghiệm người dùng (What Is Sentiment Analysis?, 2020).

Mục tiêu chính của phân tích cảm xúc là khai thác thông tin định tính trong phản hồi của khách hàng để:

- Đánh giá mức độ hài lòng và mức độ ủng hộ của người tiêu dùng,
- Phân tích hiệu quả chiến dịch truyền thông,
- Phát hiện sớm những vấn đề tiềm ẩn trong dịch vụ hoặc sản phẩm.

Các loại phân tích cảm xúc phổ biến:

- **Phân tích mức độ cảm xúc:** Phân loại cảm xúc theo nhiều cấp độ như rất tích cực, tích cực, trung lập, tiêu cực, rất tiêu cực, hoặc chấm điểm theo thang đo (ví dụ: 1–5 sao).
- **Phát hiện cảm xúc cụ thể:** Xác định cảm xúc cụ thể như vui mừng, tức giận, thất vọng, ngạc nhiên,...
- **Phân tích cảm xúc theo khía cạnh (Aspect-Based Sentiment Analysis):** Đánh giá cảm xúc đối với từng thành phần cụ thể của sản phẩm hoặc dịch vụ, chẳng hạn như chất lượng, giá cả, thái độ phục vụ.

Các phương pháp tiếp cận:

- **Dựa trên luật (Rule-based):** Sử dụng từ điển cảm xúc và tập hợp các quy tắc ngôn ngữ học để xác định tính tích cực hoặc tiêu cực của văn bản.
- **Dựa trên học máy (Machine Learning-based):** Sử dụng các thuật toán như Support Vector Machine (SVM), Naïve Bayes, Decision Tree... để huấn luyện mô hình phân loại cảm xúc dựa trên dữ liệu đã gán nhãn.
- **Dựa trên học sâu (Deep Learning-based):** Áp dụng mạng nơ-ron sâu (RNN, LSTM, Transformer) để phân tích cảm xúc với khả năng học biểu diễn ngữ nghĩa phức tạp, đạt độ chính xác cao hơn trong các tình huống đa ngữ cảnh.

4. API Gemini

Gemini API là giao diện lập trình ứng dụng do Google phát triển, cho phép người dùng truy cập và tương tác với các mô hình ngôn ngữ sinh (generative models) tiên tiến thuộc dòng Gemini. Các mô hình này được thiết kế để xử lý và tạo ra nội dung sáng tạo từ nhiều loại dữ liệu đầu vào như văn bản, hình ảnh, âm thanh và mã nguồn (Gemini API Overview | Google AI for Developers, n.d.).

Điểm nổi bật của Gemini API là tính dễ sử dụng và khả năng tiếp cận cao, ngay cả đối với những người không có nền tảng chuyên sâu về học máy. Người dùng chỉ cần mô tả nhiệm vụ đầu ra bằng ngôn ngữ tự nhiên, hệ thống sẽ tự động xử lý, mà không cần thu thập dữ liệu hoặc huấn luyện mô hình từ đầu.

Gemini API cung cấp nhiều chức năng mạnh mẽ, bao gồm:

- Tạo nội dung sáng tạo (creative content generation);
- Dịch thuật ngôn ngữ (language translation);

- Trích xuất và tóm tắt thông tin (information extraction and summarization);
- Viết và hỗ trợ lập trình (code generation and completion);
- Chuyển đổi hoặc phân tích dữ liệu phức tạp.

Nhờ vào sự linh hoạt và khả năng tổng quát hóa cao của mô hình, Gemini API trở thành công cụ hữu ích trong các ứng dụng thực tiễn như trợ lý ảo, chatbot, giáo dục, phân tích tài liệu, và hỗ trợ phát triển phần mềm (About Generative Models | Google AI for Developers, n.d.).

5. Thuật toán

5.1 Logistic Regression

Hồi quy Logistic là một thuật toán học có giám sát (supervised learning) được sử dụng phổ biến trong các bài toán phân loại, đặc biệt là phân loại nhị phân. Mục tiêu của mô hình là ước lượng xác suất một quan sát thuộc về một lớp cụ thể thông qua hàm sigmoid:

$$P(y = 1 | x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Trong đó:

- x là vector đầu vào;
- w là vector trọng số (weights);
- b là hệ số điều chỉnh (bias);
- $P(y=1|x)$ là xác suất quan sát thuộc về lớp dương

Hàm mất mát thường dùng là cross-entropy, giúp tối ưu hóa độ chính xác phân loại. Logistic Regression có ưu điểm là dễ triển khai, thời gian huấn luyện nhanh, không yêu cầu cấu trúc dữ liệu phức tạp, và hiệu quả trên

các tập dữ liệu nhỏ hoặc trung bình. Tuy nhiên, mô hình này bị giới hạn trong việc học quan hệ phi tuyến.

5.2 SVM (Super Vector Machine)

SVM là thuật toán học có giám sát dùng để giải bài toán phân loại và hồi quy. Ý tưởng chính là tìm siêu phẳng (hyperplane) tối ưu phân tách các lớp trong không gian đặc trưng, đồng thời tối đa hóa khoảng cách (margin) giữa các lớp.

Nếu dữ liệu không tuyến tính, SVM sử dụng kỹ thuật kernel trick để ánh xạ dữ liệu vào không gian có chiều cao hơn, nơi có thể tìm được siêu phẳng phân tách.

Hàm mục tiêu:

$$\min \frac{1}{2} \|\omega\|^2 \text{ với ràng buộc } y_i(\omega^T x_i + b) \geq 1$$

Ưu điểm:

- Hiệu quả với dữ liệu có chiều cao (high-dimensional);
- Khả năng khái quát hóa tốt;
- Phù hợp với dữ liệu phân lớp rõ ràng.

Nhược điểm là thời gian huấn luyện lâu khi dữ liệu lớn và khó điều chỉnh nếu dữ liệu nhiễu.

5.3 Random Forest

Random Forest là mô hình học có giám sát thuộc nhóm học tổ hợp (ensemble learning), kết hợp nhiều cây quyết định (decision trees) để cải thiện độ chính xác và giảm hiện tượng quá khớp (overfitting). Mỗi cây

trong rừng được huấn luyện trên tập con dữ liệu ngẫu nhiên (bagging) và khi dự đoán, kết quả sẽ được lấy theo hình thức bỏ phiếu đa số (majority voting) đối với phân loại hoặc trung bình hóa (averaging) với hồi quy.

Thuật toán mang lại hiệu suất cao, linh hoạt với cả dữ liệu có cấu trúc và phi cấu trúc, ít nhạy cảm với nhiễu và không yêu cầu chuẩn hóa dữ liệu.

Tuy nhiên, mô hình có thể thiếu minh bạch và khó diễn giải so với các mô hình tuyến tính.

5.4 ANN (Artificial Neuron Network)

Mạng nơ-ron nhân tạo mô phỏng hoạt động của nơ-ron sinh học. Một mạng nơ-ron đơn giản bao gồm ba thành phần chính: lớp đầu vào (input layer), lớp ẩn (hidden layer), và lớp đầu ra (output layer). Tại mỗi nút (node), đầu vào được tính toán thông qua hàm kích hoạt (activation function), thường là ReLU hoặc sigmoid:

$$a = f(\omega^T x + b)$$

Trong xử lý ngôn ngữ tự nhiên, các mạng nơ-ron thường được kết hợp với Embedding Layer, giúp ánh xạ các từ thành vector có nghĩa ngữ cảnh, hỗ trợ mô hình học tốt hơn về ngữ nghĩa.

Ưu điểm:

- Có khả năng học các quan hệ phi tuyến phức tạp;
- Phù hợp với các bài toán có lượng dữ liệu lớn;
- Có thể mở rộng sang các mô hình phức tạp hơn như RNN, LSTM.

Nhược điểm là yêu cầu nhiều tài nguyên tính toán và dễ bị quá khớp nếu không có kỹ thuật regularization như dropout.

5.5 Word2Vec

Word2Vec là kỹ thuật học biểu diễn từ (word embedding) giúp ánh xạ từ ngữ sang các vector số trong không gian liên tục sao cho từ có ngữ nghĩa tương đồng sẽ gần nhau trong không gian này. Có hai kiến trúc chính:

- CBOW (Continuous Bag of Words): dự đoán từ trung tâm dựa vào ngữ cảnh xung quanh;
- Skip-gram: dự đoán ngữ cảnh xung quanh từ trung tâm.

Ưu điểm:

- Bảo toàn được ngữ nghĩa và mối quan hệ giữa các từ (king – man + woman \approx queen);
- Giúp cải thiện chất lượng của các mô hình học máy áp dụng trên văn bản.

Word2Vec thường được huấn luyện trước (pre-trained) và sử dụng như lớp đầu vào trong các mô hình NLP.

5.6 K-Fold Cross Validation

K-Fold Cross Validation là một kỹ thuật đánh giá mô hình phổ biến trong học máy nhằm kiểm định độ ổn định và khả năng tổng quát hóa. Tập dữ liệu được chia thành K phần bằng nhau. Ở mỗi vòng lặp, một phần được dùng làm tập kiểm tra (validation), còn lại là tập huấn luyện. Kết quả cuối cùng là trung bình hiệu suất của K lần huấn luyện.

Ưu điểm:

- Giảm sai lệch do phân chia dữ liệu;
- Đánh giá toàn diện hơn so với chia train/test thông thường;

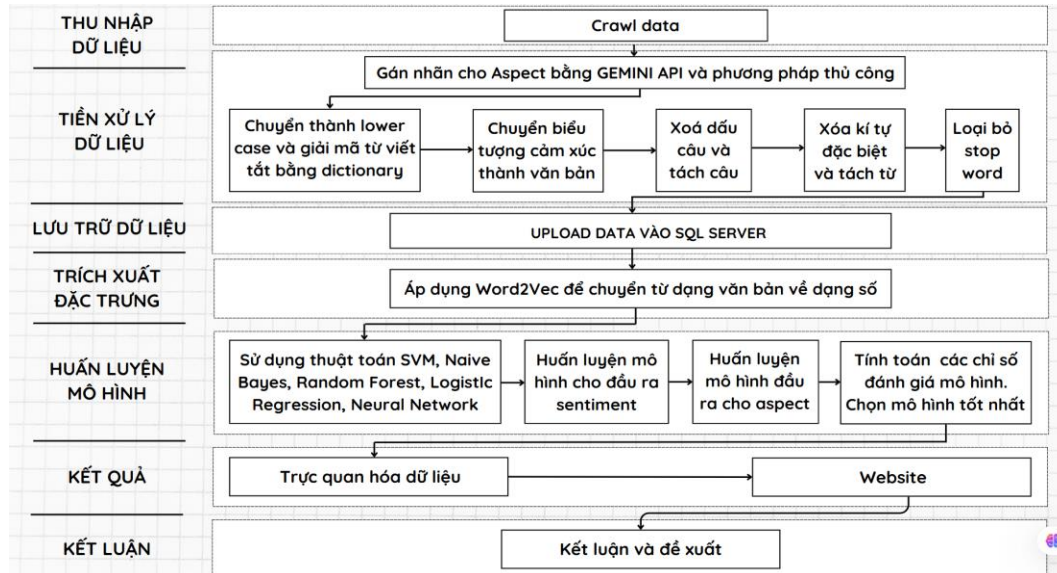
- Đặc biệt hữu ích khi dữ liệu ít.

6. Nền tảng Streamlit

Streamlit là một thư viện mã nguồn mở giúp các kỹ sư học máy xây dựng ứng dụng web giúp tương tác với người dùng (Nápoles-Duarte et al., 2022). Với thư viện này giúp cho phép thực hiện việc đưa các mô hình học máy vào trong thực tế. Streamlit giúp tiết kiệm thời gian triển khai mô hình lên ứng dụng web do có tính đơn giản, dễ tiếp cận của nó. Thư viện có khả năng triển khai các ứng dụng một cách nhanh chóng thay vì mất nhiều thời gian, nguồn lực, cũng như kiến thức về mặt kỹ thuật cần thiết để xây dựng một trang web. Chỉ với vài dòng mã Python, Streamlit có thể xây dựng nên một ứng dụng web, và không cần phải tốn thời gian cho việc lập trình front-end (Nápoles-Duarte et al., 2022). Ngoài ra, Streamlit tương tác cực kỳ tốt với các thư viện liên quan đến việc phân tích dữ liệu như Pandas, Matplotlib, và Plotly. Điều này giúp người dùng có thể dễ dàng xây dựng các bảng dashboard để trực quan hóa dữ liệu, kết hợp các mô hình dữ liệu và biểu đồ để phân tích một cách dễ dàng. Với Streamlit, thư viện không chỉ giúp hiển thị dữ liệu ở dạng tĩnh, mà còn tương thích với dữ liệu động, và trong thời gian thực.

D. PHƯƠNG PHÁP NGHIÊN CỨU

1. Cách thức triển khai



Hình 2: Framework bài toán

Để xây dựng mô hình phân tích cảm xúc và khía cạnh từ các đánh giá sản phẩm trên nền tảng thương mại điện tử Hasaki, nghiên cứu được triển khai qua một chuỗi quy trình có hệ thống và tuần tự. Trước hết, dữ liệu được thu thập từ website Hasaki thông qua công cụ web crawling kết hợp thư viện Selenium, với mục tiêu lấy được nội dung đánh giá sản phẩm của người dùng. Một số thách thức như trùng lặp đánh giá do cấu trúc website cần được xử lý bằng các hàm kiểm tra và lọc dữ liệu. Sau khi thu thập, dữ liệu được gán nhãn cho hai biến mục tiêu: sentiment (cảm xúc) và aspect (khía cạnh). Trong đó, sentiment được gán nhãn bán tự động thông qua mô hình sinh ngôn ngữ Gemini API – công cụ tiên tiến của Google có khả năng hiểu và phân tích văn bản tự nhiên. Đối với aspect, do yêu cầu đặc thù theo từng sản phẩm và lĩnh vực, nghiên cứu sử dụng phương pháp gán nhãn thủ công kết hợp với từ khóa hỗ trợ để đảm bảo tính chính xác.

Tiếp theo, dữ liệu trải qua quá trình tiền xử lý nhằm làm sạch và chuẩn hóa. Quá trình này bao gồm các bước như chuyển toàn bộ văn bản thành chữ thường, giải mã từ viết tắt, dịch thuật nếu cần, chuyển biểu tượng cảm xúc thành dạng văn bản, loại bỏ ký tự đặc biệt, tách từ, loại bỏ stop words và lemmatization. Sau khi làm sạch, dữ liệu được lưu trữ trong hệ quản trị cơ sở dữ liệu SQL Server để thuận tiện cho việc truy xuất và huấn luyện mô hình sau này. Giai đoạn tiếp theo là trích xuất đặc trưng, trong đó nghiên cứu áp dụng kỹ thuật Word2Vec để ánh xạ các từ trong văn bản thành các vector có ngữ nghĩa, giúp cải thiện độ chính xác khi huấn luyện mô hình.

Sau khi có dữ liệu biểu diễn dạng số, nghiên cứu tiến hành xây dựng và huấn luyện hai mô hình riêng biệt: một mô hình dự đoán cảm xúc (sentiment), và một mô hình dự đoán khía cạnh (aspect). Các thuật toán học máy được sử dụng trong quá trình này gồm có: Logistic Regression, Support Vector Machine (SVM), Random Forest, Naïve Bayes, và mạng nơ-ron nhân tạo (ANN). Mỗi mô hình đều được đánh giá dựa trên các chỉ số như độ chính xác (accuracy), độ nhạy (recall), độ chính xác theo dự đoán (precision) và F1-score. Mô hình tối ưu sẽ được chọn và lưu trữ phục vụ cho bước triển khai ứng dụng.

Ở giai đoạn cuối cùng, một web app được xây dựng sử dụng thư viện Streamlit, với chức năng thu thập dữ liệu mới từ sản phẩm, thực hiện tiền xử lý, áp dụng mô hình dự đoán và trực quan hóa kết quả phân tích qua dashboard. Điều này giúp người vận hành có thể theo dõi các xu hướng cảm xúc của khách hàng theo thời gian thực và đưa ra những quyết định điều chỉnh phù hợp trong kinh doanh. Từ các kết quả trực quan hóa, nghiên cứu có thể rút ra được các chủ đề nổi bật trong đánh giá khách hàng, xu hướng tích cực/tiêu cực theo dòng sản phẩm, từ đó đề xuất

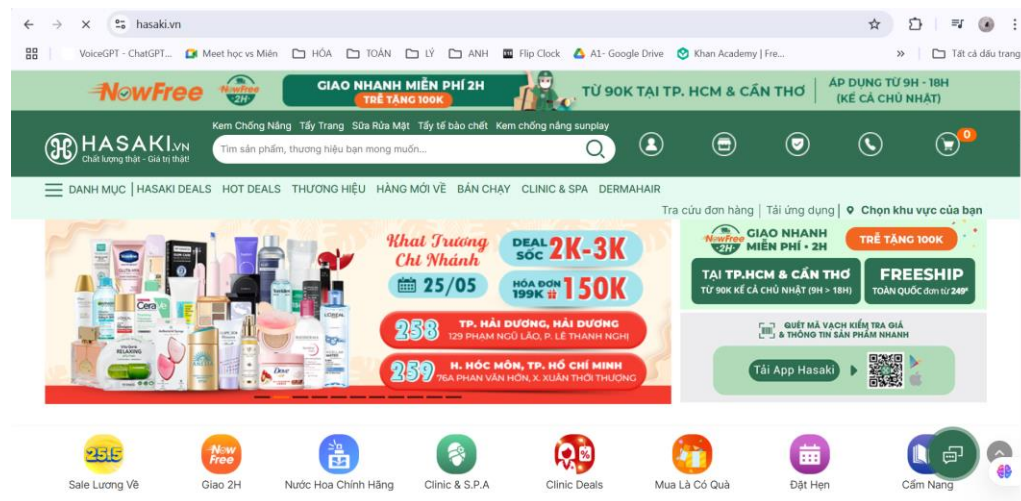
chiến lược cải thiện chất lượng dịch vụ cho Hasaki một cách kịp thời và chính xác.

2. Chi tiết từng module

2.1 Module Thu thập và Tiền xử lý dữ liệu

2.1.1. Thu thập dữ liệu

Đối với giai đoạn thu thập dữ liệu, thì nguồn dữ liệu được thu thập là từ trang web <https://hasaki.vn/>. Đây là trang thương mại điện tử của cửa hàng Hasaki chuyên bán các 21 loại mỹ phẩm, sản phẩm chăm sóc sắc đẹp, là nơi tương tác giữa khách hàng và người bán.



Hình 3: Giao diện trang thương mại điện tử Hasaki

Việc thu thập dữ liệu là trong những bước ban đầu của quá trình nghiên cứu, nên có tầm quan trọng lớn đối với kết quả. Do đó, người nghiên cứu cần thu thập dữ liệu một cách cẩn thận, cần phải xác định các loại thông tin mà nghiên cứu cần thu thập.

Đối với thông tin về sản phẩm, nghiên cứu sẽ tiến hành thu thập thông tin sản phẩm theo từng danh mục bằng cách cho duyệt vào từng danh mục sản

phẩm để thu thập thông tin. Các thông tin về sản phẩm như tên sản phẩm, giá ban đầu, giá hiện tại, số lượng đánh giá, số lượng mua hàng, tỉ lệ đánh giá đều là các thông tin mà nghiên cứu có thể thu thập. Các sản phẩm hầu hết đều được hiển thị dưới dạng bảng lưới nên tạo điều kiện rất thuận lợi cho công việc thu thập dữ liệu. Nghiên cứu cũng tiến hành thu thập các danh mục sản phẩm như:

- Chăm sóc da mặt
- Trang điểm
- Chăm sóc tóc và da đầu
- Chăm sóc cơ thể
- Nước hoa
- Chăm sóc cá nhân

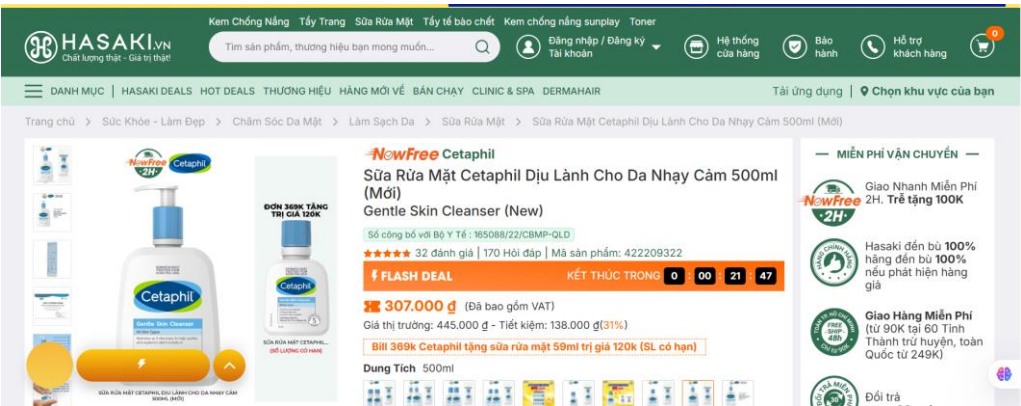


Hình 4: Trang thông tin về sản phẩm

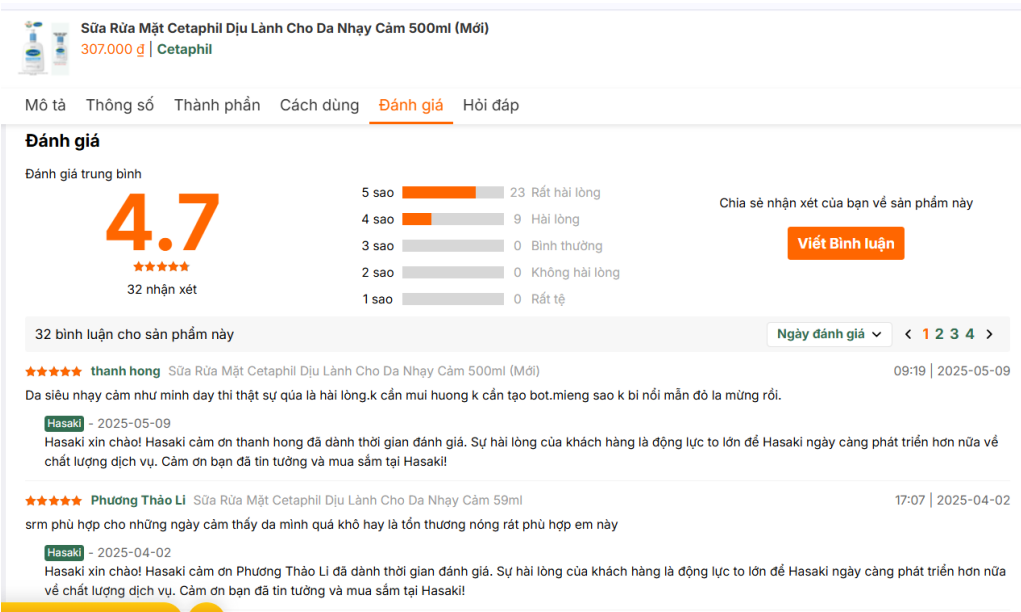
Ngoài việc thu thập thông tin về sản phẩm, nghiên cứu cũng cần thực hiện việc thu thập đối với dữ liệu đánh giá của khách hàng về sản phẩm. Để có thể làm được công việc này thì cần phải truy cập vào đường dẫn đến trang thông tin chi tiết của sản phẩm. Trang thông tin chi tiết về sản phẩm chứa các thông tin cần thu thập như tên người đánh giá, điểm đánh giá, nội dung

đánh giá, ... Các thông tin về hầu hết cũng được biểu diễn dưới dạng bảng, lưới nên rất dễ dàng cho việc thu thập.

Ngoài ra, trên trang thông tin sản phẩm cũng có rất nhiều trang đánh giá, nên nghiên cứu cần phải xây dựng hàm để có thể thu thập đến hết. Các thông tin sau khi được thu thập cần phải được kiểm tra, so khớp nhằm tránh sai sót trong quá trình thu thập.

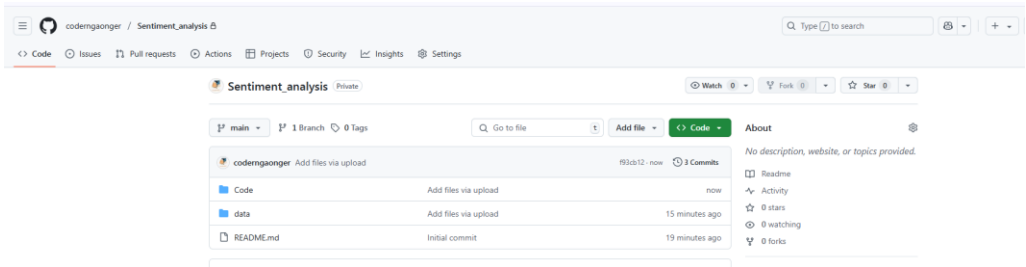


Hình 5: Trang thông tin chi tiết về sản phẩm



Hình 6: Phần đánh giá về sản phẩm của khách hàng

Đề thu thập dữ liệu từ trang thương mại điện tử Hasaki, nghiên cứu đã sử dụng thư viện Selenium và nền tảng quản lý dự án, code Github để tự động hóa quá trình thu thập, tăng tốc độ thu thập.



Hình 7: Repo của dự án

Quá trình thu thập dữ liệu là một trong những giai đoạn tốn kém nguồn lực nhất, đòi hỏi phải thực hiện một cách cẩn thận, có kiểm soát, nên việc ứng dụng công nghệ quản lý dự án Github là điều rất cần thiết. Github là hệ thống quản lý phiên bản (version control system) dựa trên nền tảng git. Thông qua Github, các dữ liệu sẽ được thu thập về có tính lịch sử dựa vào việc sử dụng các commit. Các tệp dữ liệu thu thập về sẽ được đánh dấu về mặt lịch sử. Khi có sai sót trong quá trình thực hiện, thì người dùng có thể quay lại các phiên bản cũ hơn để thực hiện lại quá trình thu thập. Ngoài ra, mỗi lần các thông tin về sản phẩm và đánh giá được thu thập về thì chúng sẽ được lưu theo từng tệp có định dạng theo ngày, tháng, năm tránh sự trùng lặp cho quá trình thu thập tiếp theo.

comment_data_20241001_1814.csv	Add files via upload	15 minutes ago
comment_data_20241001_2125.csv	Add files via upload	15 minutes ago
comment_data_20241001_2137.csv	Add files via upload	15 minutes ago
comment_data_20241001_2148.csv	Add files via upload	15 minutes ago
comment_data_20241001_2238.csv	Add files via upload	15 minutes ago
comment_data_20241001_2251.csv	Add files via upload	15 minutes ago

Hình 8: Dữ liệu thu thập được lưu theo tên có định dạng thời gian

Ngoài ra, do việc thu thập dữ liệu bằng chỉ bằng một máy tính sẽ khiến việc thu thập diễn ra tốn kém rất nhiều về mặt thời gian cũng như các

nguồn lực khác, nên nghiên cứu đã sử dụng số lượng máy tính nhiều hơn để tăng tốc cho việc thu thập. Tuy nhiên, điều này sẽ quá trình thu thập khác nhau nên dữ liệu hai máy sẽ dẫn đến khác nhau. Do vậy, nghiên cứu ứng dụng Github để giải quyết vấn đề này, để giúp các máy có thể hợp nhất về mặt lịch sử, giúp công việc thu thập trên hai hoặc cả là nhiều máy diễn ra thuận lợi hơn. Đối với từng máy, một nhánh mới sẽ được tạo ra, và sau khi quá trình thu thập diễn ra trên nhánh đó, thì nghiên cứu sẽ cập nhật nhánh chính về trạng thái mới nhất, và hợp nhất commit trên nhánh mới đó và nhánh chính.

Về quá trình thu thập dữ liệu thì quá trình này được chia thành hai giai đoạn. Trong giai đoạn đầu tiên, nghiên cứu sẽ thu thập thông tin về các sản phẩm được bán trên trang thương mại điện tử Hasaki. Cụ thể là người dùng cần phải xác định đường dẫn của danh mục sản phẩm cần thu thập. Một đường dẫn như vậy có rất nhiều trang. Việc mà cần làm là duyệt tự động vào từng trang này để thu thập dữ liệu cung cấp thông tin chi tiết về các sản phẩm: mã sản phẩm, mã sản phẩm trong cơ sở dữ liệu của Hasaki, đường dẫn truy cập của sản phẩm, tiêu đề sản phẩm, giá gốc và giá hiện tại, số lượng sản phẩm đã bán ra, biến thể sản phẩm, và số lượng đánh giá từ khách hàng. Dữ liệu mà giai đoạn thu thập được sẽ được lưu trong bộ nhớ dưới dạng file csv.

A	B	C	D	E	F	G	H	I
DataProd	LinkItem	DataProd	ProductVariant	CurrentPri	QuantitySi	TotalRevenue	TotalQuantity	
845	https://ha['843', '84Xit Khoáng Avène Cấp Nước, Làm Dịu & Giảm Kích Ứng 300ml			318000	315	100170000	315	
983	https://ha['83427', '5Mặt Nạ Ngủ Laneige Cung Cấp Nước 70ml			570000	48	54720000	96	
987	https://ha['995', '11Xit Khoáng La Roche-Posay Làm Dịu Và Bảo Vệ Da 50g			99000	146	14454000	146	
988	https://ha['988', '40Xit Khoáng Vichy Làm Dịu, Củng Cố & Cấp Ẩm Cho Da 300ml			364000	26	9464000	26	
990	https://ha['4041', '9Xit Khoáng Vichy Làm Dịu, Củng Cố & Cấp Ẩm Cho Da 300ml			144000	29	4176000	29	
1001	https://ha['987', '99Xit Khoáng La Roche-Posay Làm Dịu Và Bảo Vệ Da 50g			379000	98	37142000	98	
1278	https://ha['1279', '1Xit Khoáng Evoluderm Dưỡng Ẩm Cấp Nước & Làm Dịu Da 400ml			139000	679	94381000	679	
1279	https://ha['61390', 'Xit Khoáng Evoluderm Dưỡng Ẩm Cấp Nước & Làm Dịu Da 400ml			99000	154	15246000	154	
1294	https://ha['1424', '1Sữa Tẩy Trang Hằng Ngày Dịu Nhẹ - 250ml			89000	11	979000	11	
1343	https://ha['14949', 'tNước Hoa Hồng Evoluderm Cho Da Hỗn Hợp Và Dầu Mụn 250ml			99000	33	3267000	33	
1734	https://ha['1734', '1Lotion Chấm Mụn Bye Bye Blemish Dành Cho Mụn Sưng Viêm 30ml			297000	4	1188000	4	
1735	https://ha['94457', 'Nước Tẩy Trang Bioderma Dành Cho Da Nhạy Cảm 500ml			168000	180	30240000	180	
1738	https://ha['2625', '1Nước Tẩy Trang Bioderma Dành Cho Da Dầu & Hỗn Hợp 500ml			168000	137	23016000	137	
1746	https://ha['108742', 'Kem Dưỡng Bioderma Giúp Se Khít Lỗ Chân Lông 30ml			369000	260	95940000	260	
1910	https://ha['1910', '1Kem Chống Nắng Ultra Sheer SPF 50 88ml			311000	72	22392000	72	
1944	https://ha['88835', 'Kem Chống Nắng Vichy Ngừa Sạm & Thâm Nám (Màu Da) 50ml			468000	37	17316000	37	
1971	https://ha['1972', '9XSữa Rửa Mặt St.Ives Tẩy Tế Bào Chết Hoa Hồng & Lô Hội 170g			109000	227	24743000	227	

Hình 9: Dữ liệu chứa thông tin về sản phẩm thu thập được từ trang Hasaki

Tên cột	Kiểu dữ liệu	Giá trị	Mô tả
product_id	int	422213231	Định danh duy nhất của sản phẩm trên Hasaki
data_product_id	int	107938	Mã sản phẩm trong cơ sở dữ liệu của Hasaki, có thể là một mã sản phẩm nội bộ
link_item	text	https://hasaki.vn/san-pham/combo-2-bong-tay-trang-hotosu-cao-cap-150-mieng-107938.html	Đường dẫn đến trang sản phẩm trên Hasaki, cho phép truy cập trực tiếp đến thông tin chi tiết về sản phẩm
title	text	Combo 2 Bông	Tiêu đề hoặc tên của

		Tây Trang Hotosu Cao Cấp 150 Miếng	sản phẩm, cung cấp một cái nhìn tổng quan về nội dung hoặc chức năng của sản phẩm
original_price	int	70000	Giá gốc của sản phẩm trước khi có bất kỳ chiết khấu hoặc khuyến mãi nào
current_price	int	51000	Giá hiện tại của sản phẩm sau khi áp dụng mọi chiết khấu hoặc khuyến mãi
quantity_sold	int	3053	Số lượng sản phẩm đã bán được đến thời điểm hiện tại, cung cấp thông tin về mức độ phổ biến của sản phẩm
product_variant	text	300 miếng	Thông tin về biến thể hoặc phiên bản của sản phẩm
review_count	int	29	Số lượng đánh giá từ khách hàng đã mua hàng, cung cấp một

			chỉ số về mức độ tương tác của khách hàng với sản phẩm
--	--	--	--

Bảng 1: Thông tin dữ liệu về sản phẩm

Giai đoạn thứ hai tập trung vào việc thu thập dữ liệu đánh giá của khách hàng về các sản phẩm trên Hasaki. Tương tự như giai đoạn trước, nghiên cứu đã duyệt qua từng trang 28 đánh giá của sản phẩm và lấy thông tin về đánh giá từ khách hàng, bao gồm nội dung đánh giá, điểm đánh giá, tên người đăng, và thời gian đăng. Khi thực hiện công việc này, nghiên cứu đồng thời cũng thu thập các thông tin đi kèm như: các đánh giá đường dẫn đến trang sản phẩm, danh sách mã các biến thể của sản phẩm, biến thể sản phẩm, mã sản phẩm, tên khách hàng đánh giá, nội dung đánh giá, thời gian đăng và điểm đánh giá của người dùng. Dữ liệu sẽ rất có ích cho bài toán, giúp doanh nghiệp hiểu rõ hơn về ý kiến và cảm xúc của khách hàng đối với từng sản phẩm. Tuy nhiên, ở giai đoạn này, do nhiều sản phẩm có chung hệ thống đánh giá với nhau, nên khiến dữ liệu bình luận thu được bị trùng lặp. Do đó nghiên cứu cần phát triển hàm, phương pháp để loại bỏ sự trùng lặp và quyết định liệu thu thập hay không. Điều này sẽ giúp tiết kiệm thời gian thu thập, và giúp thu lại bộ dữ liệu tốt nhất. Dữ liệu này ở giai đoạn hai sẽ được kết hợp với dữ liệu từ công đoạn trước tạo ra một bộ dữ liệu hoàn chỉnh về đánh giá của khách hàng đối với từng sản phẩm, phục vụ cho công việc phân tích.

Tên cột	Kiểu dữ liệu	Giá trị	Mô tả
link_item	text	https://hasaki.vn/sa-n-pham/thuoc-nhuom-toc-phe-bo-dang-kem-bigen-882-nau-den-98129.html	Đường dẫn đến trang sản phẩm trên Hasaki, cho phép truy cập trực tiếp đến thông tin chi tiết về sản phẩm và sản phẩm liên quan đến bình luận
data_product_id_list	list	["14397", "14379", "14389", "14307", "98129", "98127", "14409"]	Danh sách các mã sản phẩm, chứa product_id liên quan đến bình luận
data_product_id	int	98129	Mã sản phẩm trong cơ sở dữ liệu của Hasaki, cung cấp thông tin về sản phẩm đang được đánh giá
name_comment	text	Tâm Nguyễn	Tên hoặc biệt danh của người dùng bình luận
content_comment	text	Tóc mình nhuộm màu cách đây 6 tháng, giờ mình muốn nhuộm đen này có bám tốt không ạ?	Nội dung của bình luận, giúp hiểu được phản hồi của người dùng về sản phẩm

product_variant	text	Thuốc Nhuộm Tóc Bigen Phủ Bạc Dạng Kem S82 Nâu Đen 80ml Đầu mua hàng online	Thông tin về biến thể hoặc phiên bản của sản phẩm, nếu có
datetime_comment	date	10:03	18/08/2023
rating	int	4	Số sao người dùng đánh giá cho sản phẩm (thang điểm 5), cho biết mức độ hài lòng của người dùng khi sử dụng hoặc tương tác với sản phẩm

Bảng 2: Mô tả dữ liệu chứa thông tin về đánh giá của sản phẩm

Dữ liệu sẽ được thu thập cùng lúc bằng nhiều máy khác và dần dần cho đến hết. Trong quá trình đó, dữ liệu mới thu thập sẽ được cập nhật bằng cách đẩy lên và kéo về từ Github và hợp nhất vào nhánh và lịch sử commit. Dưới sự hỗ trợ nền tảng quản lý hệ thống quản lý dự án và phiên bản code Github, dữ liệu có thể được thu thập đồng thời, nên quá trình thu thập diễn ra nhanh hơn thông qua việc cập nhật chung vào một lịch sử commit. Dữ liệu sau khi thu thập sẽ được tổ chức và lưu trữ dữ liệu trong các tập tin CSV để tiện cho việc sử dụng và phân tích sau này. Kết hợp cả hai giai đoạn này, nghiên cứu sẽ có đủ dữ liệu để thực hiện phân tích cảm xúc của khách hàng dựa trên đánh giá sản phẩm. Sự kết hợp thông tin từ cả hai tập tin dữ liệu này, kết quả thu được có thể tạo ra một cái nhìn toàn diện về sự tương tác giữa khách hàng và sản phẩm trên Hasaki, từ đó hỗ trợ quá trình phân tích và đánh giá hiệu suất của sản phẩm cũng như nhu cầu của thị trường.

2.1.2. Tiền xử lý và chuẩn hóa dữ liệu

- Gắn nhãn cho dữ liệu

Để thực hiện phân tích và gắn nhãn các bình luận của khách hàng sau khi mua mỹ phẩm trên nền tảng Hasaki, chúng tôi sử dụng Gemini API nhằm phân loại nội dung theo hai chiều: nhãn thực thể và nhãn cảm xúc. Cụ thể, mỗi bình luận trong cột “Comment” sẽ được tách thành từng câu đơn, và mỗi câu đều bắt buộc phải được gắn một nhãn thực thể như bao bì (PACKAGING), chất lượng (QUALITY), giá cả (PRICE), dịch vụ (SERVICE), cửa hàng (STORE) hoặc nội dung khác (MISCELLANEOUS), cùng với một nhãn cảm xúc tương ứng là tích cực (positive), tiêu cực (negative) hoặc trung lập (neutral). Việc gắn nhãn được thực hiện hoàn toàn tự động thông qua yêu cầu gửi đến Gemini API với định dạng đầu vào rõ ràng và chuẩn hóa. Kết quả trả về sẽ được lưu vào một cột mới có tên “Label analysis”, trong đó thể hiện chi tiết từng câu trong bình luận đã được gắn nhãn như thế nào. Sau khi nhận kết quả, chúng tôi tiến hành rà soát lại dữ liệu để kiểm tra và hiệu chỉnh các trường hợp phân tích sai nhãn nếu có. Đồng thời, thêm một cột “Comment_ID” để định danh từng phản hồi duy nhất. Cuối cùng, toàn bộ dữ liệu được lưu lại dưới dạng file CSV với tên là hasaki_data.csv để phục vụ các bước phân tích hoặc trực quan hóa tiếp theo. Ví dụ, với bình luận như “Sản phẩm đóng gói rất đẹp. Hiệu quả dưỡng ẩm rất tốt. Giá cả hợp lý. Giao hàng chậm, cần cải thiện. Bao bì chắc chắn, nhưng sản phẩm hơi đắt.”, hệ thống sẽ tự động phân tích và cho ra kết quả tương ứng: câu đầu tiên được gắn là PACKAGING với cảm xúc positive, câu thứ hai là QUALITY với positive, câu thứ ba là PRICE với positive, câu thứ tư là SERVICE với negative, câu thứ năm là PACKAGING với positive và câu cuối cùng là PRICE với negative. Cách tiếp cận này giúp

trích xuất thông tin có cấu trúc từ phản hồi khách hàng, phục vụ hiệu quả cho mục tiêu phân tích trải nghiệm và cải tiến sản phẩm, dịch vụ.

Sau đó sẽ tiến hành bóc tách nhãn thực thể và nhãn cảm xúc từ “Label analysis” và tạo ra một dataframe mới gồm 4 cột: “Comment_ID”, “Comment”, “Entity label”, “Sentiment label”.

Sau đó, lưu trữ dữ liệu dưới dạng file csv, với tên file là label.csv.

Tổng quan dữ liệu sau khi bóc tách nhãn thực thể và nhãn cảm xúc:

content_ID	content	Entity label	Sentiment label
0	Kem chống nắng không nâng tone	QUALITY	negative
0	nhưng hơi tối màu khi lên da	QUALITY	negative
1	Sữa rửa mặt này dùng rất ok	QUALITY	positive
2	Chất kem dễ tán	QUALITY	positive
2	không nâng tone	QUALITY	neutral
2	nhưng bôi lên mặt không ráo mà cứ bóng bóng	QUALITY	negative
3	Tắm xong còn thơm thoang thoang quanh người	QUALITY	positive
4	Ban cũ rất thích	MISCELLANEOUS	neutral
4	bản mới dùng bị châm chít khoảng 2phút là hết	QUALITY	negative
4	vẫn thích bản cũ hơn	MISCELLANEOUS	neutral
5	Mỗi lần da mình khô hay bị rát là đắp này thấy dịu da cấp ẩm đỉnh của chóp	QUALITY	positive

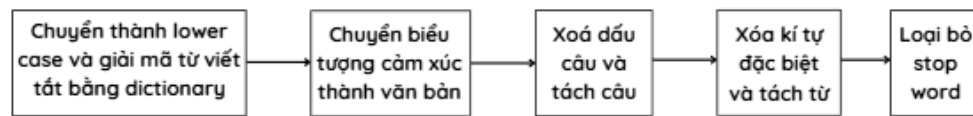
Hình 10: Tổng quan dữ liệu sau khi bóc tách nhãn thực thể

Tiếp theo đó sẽ gộp dữ liệu thành 1 comment thống nhất và chia ra loại cảm xúc cho từng Entity

content_ID	content	QUALITY	STORE	PRICE	SERVICE	PACKAGING	OTHERS
0	Sản phẩm dùng rất tốt phù hợp với da mặt mình	positive	neutral	neutral	neutral	neutral	neutral
1	Mặt nạ đắp hiệu quả thật	positive	neutral	neutral	neutral	neutral	neutral
2	Sử dụng dễ dàng rất thoải mái thư giãn tốt độ	positive	neutral	neutral	neutral	neutral	neutral
3	Sản phẩm này dùng cực kỳ êm luôn	positive	neutral	neutral	neutral	neutral	neutral
4	Sản phẩm dùng rất tốt	positive	neutral	neutral	neutral	neutral	neutral
5	Xài sạch mà không gây khô da sản phẩm rất tốt	positive	neutral	neutral	neutral	neutral	neutral
6	Dùng combo cùng serum cực kì ổn	positive	neutral	neutral	neutral	neutral	neutral
7	Sản phẩm tốt Hiệu quả	positive	neutral	neutral	neutral	neutral	neutral
8	Sử dụng dễ dàng rất thoải mái thư giãn tốt độ	positive	neutral	neutral	neutral	neutral	neutral
9	Đắp vào các nốt mụn sưng giảm hẳn sạch các sợi bã nhờn	positive	neutral	neutral	neutral	neutral	neutral
10	Sản phẩm chất lượng đáng sử dụng	positive	neutral	neutral	neutral	neutral	neutral
11	Em mới mua bên hasaki chai srm này mà mới dùng vài lần ai muốn mua về c	neutral	neutral	positive	neutral	neutral	neutral

Hình 11: Kết quả sau khi gộp comment

- Tiền xử lý dữ liệu



Hình 12: Quy trình tiền xử lý dữ liệu

Quy trình tiền xử lý dữ liệu là một bước không thể thiếu và có ý nghĩa then chốt trong việc chuẩn bị dữ liệu cho các mô hình phân tích cảm xúc. Mục tiêu của giai đoạn này là chuẩn hóa, làm sạch và tối ưu nội dung phản hồi nhằm loại bỏ các yếu tố gây nhiễu, đồng thời giữ lại những thông tin cốt lõi, góp phần nâng cao hiệu suất mô hình phân tích. Đầu tiên, văn bản sẽ được chuyển đổi toàn bộ thành chữ thường (lowercase) để đảm bảo tính đồng nhất và giúp việc dịch sang tiếng Anh diễn ra thuận lợi hơn. Kế tiếp, các từ viết tắt phổ biến hoặc mang tính chuyên ngành, chẳng hạn trong lĩnh vực mỹ phẩm, sẽ được giải mã thông qua một từ điển xây dựng sẵn. Trong nhiều trường hợp, việc dịch sang tiếng Anh bằng Google Translate cũng hỗ trợ tự động mở rộng khả năng giải nghĩa những từ viết tắt này. Sau đó, dữ liệu sẽ được dịch sang tiếng Anh, bước này không chỉ chuẩn hóa ngôn ngữ mà còn góp phần đồng nhất nội dung bình luận. Tiếp đến, các biểu tượng cảm xúc như “:)”, “:(”, “:D” sẽ được chuyển đổi thành các từ văn bản tương ứng như “happy”, “sad”, “laugh” để làm rõ sắc thái cảm xúc. Đồng thời, hệ thống sẽ loại bỏ các ký tự đặc biệt như dấu chấm, dấu phẩy, dấu ngoặc hoặc các ký hiệu không cần thiết vì đây là những yếu tố gây nhiễu, không mang giá trị phân tích. Sau khi văn bản đã được làm sạch sơ bộ, bước tokenization sẽ được thực hiện để tách văn bản thành từng từ hoặc cụm từ riêng biệt. Tiếp theo, các từ vô nghĩa (stop words) như “how”, “what”, “is”, “are” – vốn không đóng góp thông tin đáng kể – sẽ bị loại bỏ nhằm tăng mức độ tập trung vào các đặc trưng quan trọng hơn. Cuối cùng, văn bản sẽ được đưa về dạng gốc thông qua quá trình lemmatization. Việc

đưa từ về nguyên thể giúp giảm thiểu biến thể từ vựng và đảm bảo tính đồng nhất cho tập dữ liệu. Toàn bộ quy trình này sẽ góp phần tạo nên một tập dữ liệu sạch, nhất quán và tối ưu, làm tiền đề cho việc huấn luyện và phân tích bằng các mô hình học máy đạt hiệu quả cao hơn.

index	abbreviation	meaning
1	+	và
2	10đ	10 điểm
3	10đ	10 điểm
4	bb	bạn bè
5	bn	bạn
6	bông tt	bông tẩy trung
7	bth	bình thường
8	cg	cũng
9	cx	cũng
10	da hh	da hỗn hợp
11	đc	được
12	dth	dễ thương
13	é	á
14	h	giờ
15	hh	hỗn hợp
16	hh thiên dầu	hỗn hợp thiên dầu
17	hh thiên khô	hỗn hợp thiên khô
18	hhtd	hỗn hợp thiên dầu
19	hhtk	hỗn hợp thiên khô
20	hk	không

Hình 13: Từ điển giải mã một số từ viết tắt

Cuối cùng, từ dữ liệu đã được tiền xử lý, bài toán sẽ thu được kết quả như sau. Dữ liệu này sẽ được sử dụng cho giai đoạn tiếp theo là trích xuất đặc trưng.

[illegible]

Hình 14: Dữ liệu sau khi được tiền xử lý

Đồng thời nghiên cứu cũng trực quan hóa các token để có cái nhìn tổng quan về bộ dữ liệu đã được tiền xử lý. Hình ảnh trực quan từ biểu đồ cho thấy việc tiền xử lý dữ liệu diễn ra một cách thuận lợi. Các từ stopwords, ký tự đặc biệt đều đã được loại bỏ. Và các từ đều ở dạng chữ thường.



Hình 15: Word Cloud

2.2 Module Huấn luyện mô hình

Ở bước chuẩn bị dữ liệu cho quá trình huấn luyện, đây là giai đoạn mà cần gán nhãn cho các các đánh giá nằm trong bộ dữ liệu huấn luyện. Sau đó dữ liệu sẽ được đưa vào mô hình. Trong bài toán này, sentiment_label được sử dụng để gán nhãn cho cảm xúc của khách hàng trong đánh giá, trong khi aspect_label được sử dụng để gán nhãn cho các khía cạnh cụ thể của sản phẩm được đề cập trong đánh giá.

Trong phần huấn luyện mô hình phân tích cảm xúc của khách hàng dựa vào đánh giá sản phẩm bằng cách sử dụng mô hình học máy, công việc nghiên cứu làm là sẽ tiến hành huấn luyện mô hình trên tập dữ liệu đã được chuẩn bị trước đó. Mục tiêu của quá trình huấn luyện là để mô hình học được mối quan hệ giữa các đặc trưng của đánh giá sản phẩm và nhãn cảm xúc hoặc khía cạnh của sản phẩm được gán nhãn. Trước khi huấn luyện mô hình, bài toán cần chia bộ dữ liệu thành hai bộ dữ liệu huấn luyện và kiểm tra. Đối với mỗi đầu ra, thì cần phải huấn luyện một mô hình. Mô hình sau khi được huấn luyện sẽ tiếp tục dùng cho dữ liệu mới. Quá trình huấn luyện giúp mô hình học được các đặc trưng, mối quan hệ trên dữ liệu đa dạng và có khả năng tổng quát hóa tốt khi áp dụng vào dữ liệu mới

- Với mô hình Neural Network:

Áp dụng Word2Vec và các kỹ thuật khác để chuyển từ dạng văn bản về dạng số (vector):

- + Chuẩn bị dữ liệu cho mô hình Neural Network:

1. Chuyển nhãn cảm xúc Output thành giá trị số (bắt buộc cho Neural Network).

2. Tạo tokenizer cho padded sequences(biến được khai báo ở sau) có nghĩa là sẽ token(chuyển hoá) từng từ đã xử lý thành số thứ tự (thứ tự được xếp theo tần suất xuất hiện của từ đó giảm dần trong từ điển).
3. Sau đó lưu lại từ điển đã Tokenizer bằng thư viện Joblib với cú pháp “dump”.

```
# Chuyển nhãn cảm xúc thành số
label_mapping = {"positive": 2, "negative": 0, "neutral": 1}
train_labels = np.array([label_mapping[label] for label in train_labels])
test_labels = np.array([label_mapping[label] for label in test_labels])

# Tạo tokenizer cho padded sequences
tokenizer = Tokenizer()
tokenizer.fit_on_texts(train_comments)

# Lưu tokenizer
dump(tokenizer, '/content/drive/MyDrive/Final ML/tokenizer.pkl')

word_index = tokenizer.word_index
print(word_index)
```

4. Chuyển dữ liệu thành padded sequences(Vì Neural Network yêu cầu đầu vào dữ liệu phải đồng nhất về kích thước nên khi chuyển sang padded sequences sẽ giúp các dữ liệu đầu vào có cùng kích thước với nhau. Ví dụ kích thước hiện tại đang là 100 thì nếu dữ liệu đầu vào là một câu có vector số không đủ 100 chiều(từ) thì sẽ được thêm các giá trị 0 đằng sau để đủ 100 chiều hoặc các từ quá 100 chiều sẽ bị chia đôi ra

```
# Chuyển dữ liệu thành padded sequences
max_length = 100
train_sequences = tokenizer.texts_to_sequences(train_comments)
test_sequences = tokenizer.texts_to_sequences(test_comments)
print(train_sequences)
train_padded = pad_sequences(train_sequences, maxlen=max_length, padding='post', truncating='post')
test_padded = pad_sequences(test_sequences, maxlen=max_length, padding='post', truncating='post')
```

5. Chuyển nhãn thành dạng one hot vector cho Neural Network


```
# Chuyển nhãn thành dạng one-hot vector cho Neural Network
train_labels_one_hot = to_categorical(train_labels, num_classes=3)
test_labels_one_hot = to_categorical(test_labels, num_classes=3)
```

+ Chuẩn bị dữ liệu cho các mô hình còn lại:

1. Huấn luyện Word2Vec và lưu mô hình:

```
# Huấn luyện Word2Vec
processed_train_comments = [comment for comment in train_comments]
print(processed_train_comments)
w2v_model = Word2Vec(sentences=processed_train_comments, vector_size=100, window=3, min_count=1, workers=4)
# Lưu mô hình Word2Vec
w2v_model.save('/content/drive/MyDrive/Final ML/word2vec_sentiment.model')
```

2. Xây dựng hàm chuyển đổi dữ liệu sang vector đặc trưng bằng Word2Vec:

```
# Hàm chuyển đổi dữ liệu sang vector đặc trưng bằng Word2Vec
def vectorize_comment(comment, model, vector_dim):
    words = comment
    word_vectors = [model.wv[word] for word in words if word in list(model.wv.index_to_key)]
    if len(word_vectors) == 0:
        return np.zeros(vector_dim) # Nếu không có từ nào trong mô hình Word2Vec
    return np.mean(word_vectors, axis=0)

# Tạo vector đặc trưng
vector_dim = 100
train_vectors = np.array([vectorize_comment(comment, w2v_model, vector_dim) for comment in train_comments])
test_vectors = np.array([vectorize_comment(comment, w2v_model, vector_dim) for comment in test_comments])
```

- Huấn luyện mô hình
- + Mô hình Neural Network

1. Hàm sử dụng để huấn luyện:

```
# Hàm huấn luyện Neural Network
def train_neural_network(X_train_fold, y_train_fold, X_valid_fold, y_valid_fold, embedding_matrix, word_index, vector_dim):
    model = Sequential()
    model.add(Embedding(input_dim=len(word_index) + 1, output_dim=vector_dim,
                        weights=[embedding_matrix], trainable=True))
    model.add(Flatten())
    model.add(Dense(64, activation='relu'))
    model.add(Dropout(0.3))
    model.add(Dense(64, activation='relu'))
    model.add(Dropout(0.3))
    model.add(Dense(3, activation='softmax'))
    model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

    # EarlyStopping callback
    early_stopping = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True)

    model.fit(X_train_fold, y_train_fold,
              validation_data=(X_valid_fold, y_valid_fold),
              epochs=15, batch_size=64, verbose=0, callbacks=[early_stopping])
    return model
```

2. Chuẩn bị Embedding Matrix để đưa vào làm trọng số trong lớp Embedding Layer trong hàm Neural Network:

```
# Chuẩn bị embedding matrix
embedding_matrix = np.zeros((len(word_index) + 1, vector_dim))
for word, i in word_index.items():
    if word in list(w2v_model.wv.index_to_key):
        embedding_matrix[i] = w2v_model.wv[word]
print(embedding_matrix)
```

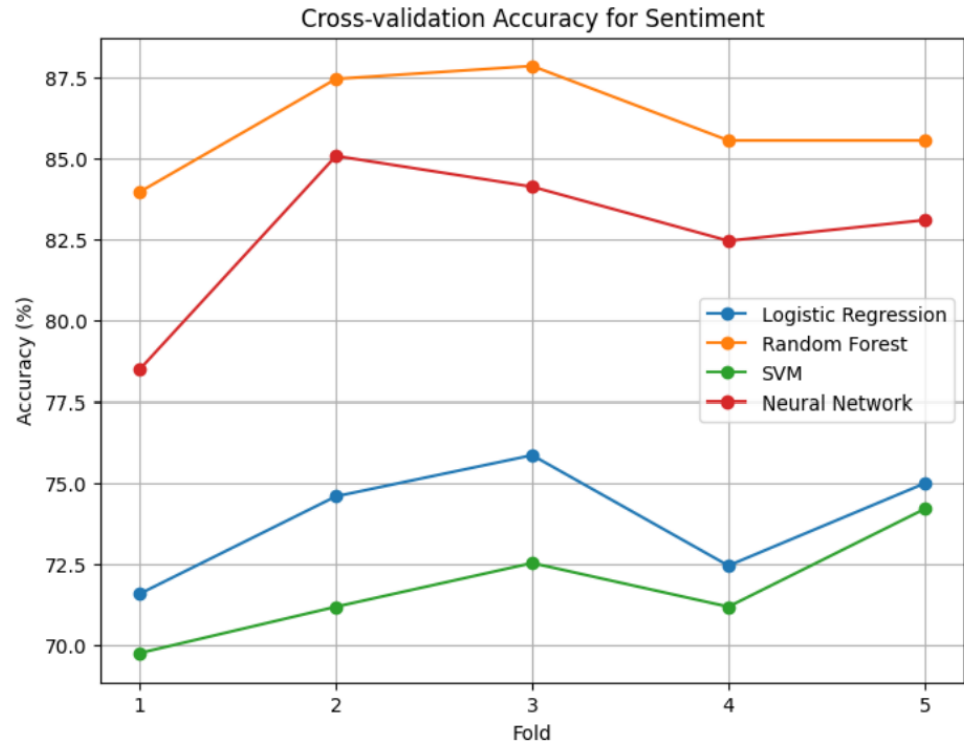
3. Hàm đánh giá độ chính xác mô hình Neural Network:

```
#Đánh giá mô hình Neural Network
def evaluate_model(model, X_valid_fold, y_valid_fold):
    _, nn_accuracy = model.evaluate(X_valid_fold, y_valid_fold, verbose=0)
    return nn_accuracy * 100
```

- Áp dụng kỹ thuật K-Fold-Cross-Validation để huấn luyện và đánh giá độ chính xác của 4 mô hình trên bộ dữ liệu:

Kết quả phân loại với K-Fold Cross-Validation:						
	Fold	Logistic Regression	SVM	Random Forest	Neural Network	
0	1	71.59	69.76	83.97	78.49	
1	2	74.60	71.19	87.46	85.08	
2	3	75.87	72.54	87.86	84.13	
3	4	72.46	71.19	85.56	82.46	
4	5	75.00	74.21	85.56	83.10	

Hình 16: Kết quả của từng Fold



Hình 17: Kết quả của từng Fold trực quan

+ Tính trung bình độ chính xác và chọn ra mô hình tốt nhất:

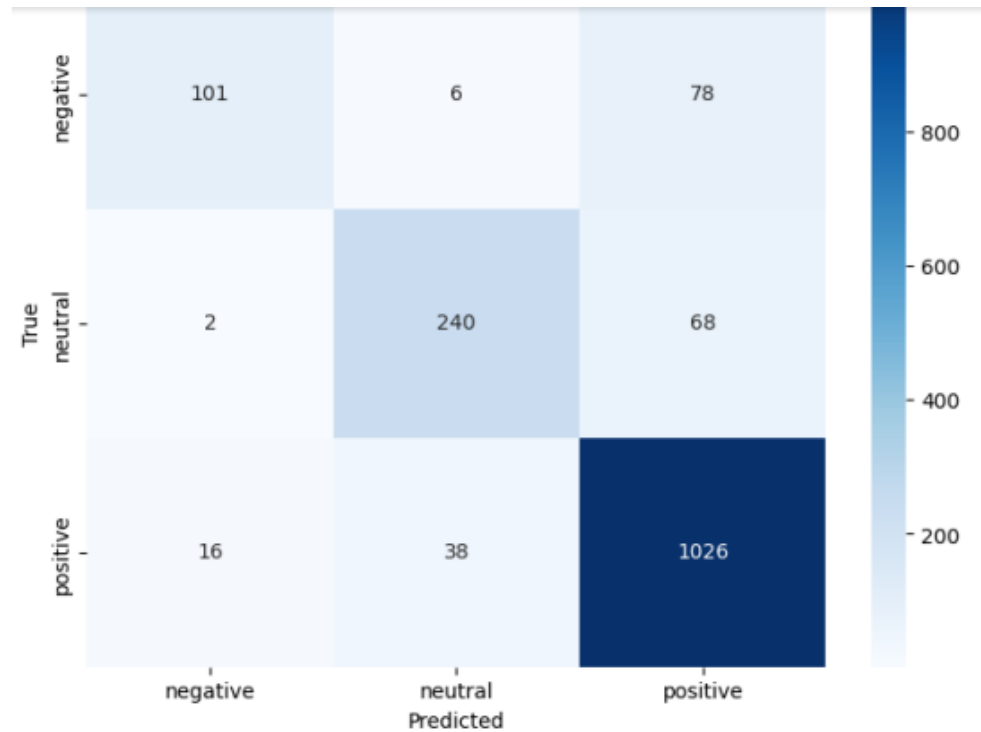
```

Mean_Accuracy for sentiment:
      Model Name  Mean_Accuracy
0  Logistic Regression      73.90
1    Random Forest      86.08
2              SVM      71.78
3    Neural Network      82.65

The best model is 'Random Forest' with a mean accuracy of 86.08.
  
```

Hình 18: Tính trung bình độ chính xác và chọn ra mô hình tốt nhất

- Đánh giá mô hình
 - + Confusion Matrix



Hình 19: Confusion Matrix

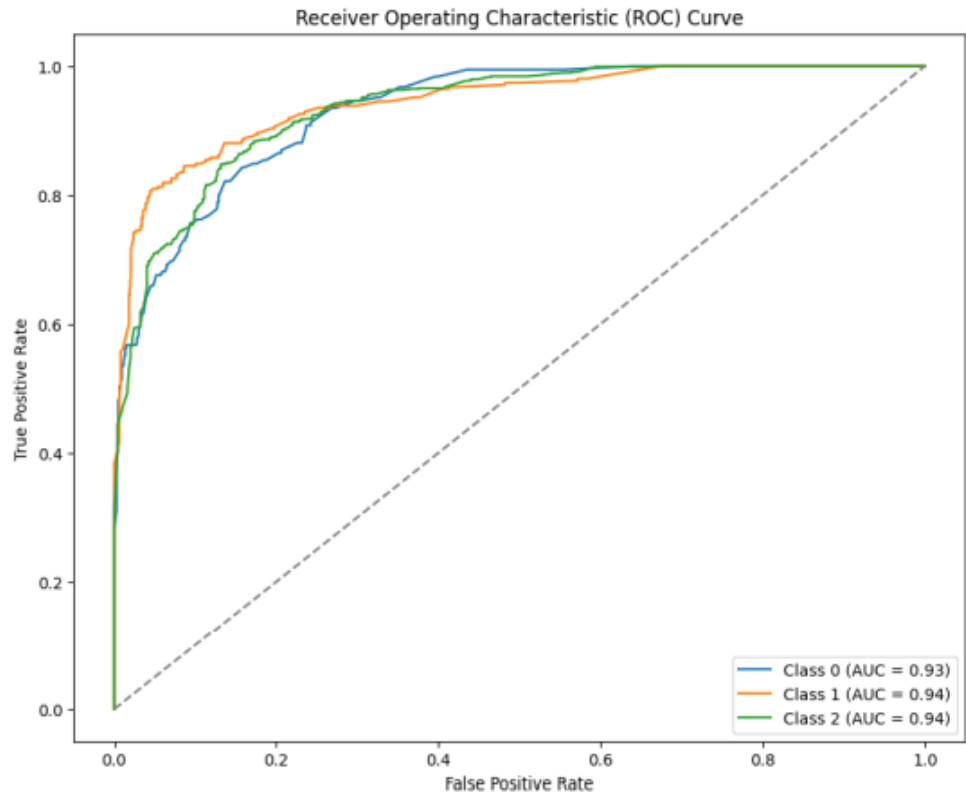
Có thể thấy mô hình phân biệt khá tốt giữa các cảm xúc khác nhau chỉ có negative có sự nhầm lẫn cao có lẽ là do dữ liệu có quá nhiều positive nên sẽ tạo ra hiện tượng thiên vị dữ liệu làm cho máy dự đoán sai

+ Classification Report

Classification Report:				
	precision	recall	f1-score	support
negative	0.85	0.55	0.66	185
neutral	0.85	0.77	0.81	310
positive	0.88	0.95	0.91	1080
accuracy			0.87	1575
macro avg	0.86	0.76	0.79	1575
weighted avg	0.87	0.87	0.86	1575

Hình 20: Classification Report

+ ROC và AUC



Hình 21: ROC và AUC

- Thực hiện phân tích cảm xúc từ 1 comment của khách hàng bất kỳ theo nhiều khía cạnh khác nhau:

```
# Ví dụ
input_text = "Hiệu quả dưỡng ẩm rất tốt, Giá cả hợp lý , Giao hàng chậm"
QUALITY, SERVICE, OTHERS, STORE, PACKAGING, PRICE = predict_sentiment(input_text)
print(f"Input: {input_text}")
print(f"QUALITY: {QUALITY}")
print(f"SERVICE: {SERVICE}")
print(f"OTHERS: {OTHERS}")
print(f"STORE: {STORE}")
print(f"PACKAGING: {PACKAGING}")
print(f"PRICE: {PRICE}")
```

Hình 22: Ví dụ

QUALITY: positive
SERVICE: negative
OTHERS: neutral
STORE: neutral
PACKAGING: neutral
PRICE: positive

Hình 23: Kết quả ví dụ

2.3 Module Website



Hình 24: Quy trình xây dựng website

Ở phần này, nghiên cứu sẽ cần đến web app để đưa bài toán, mô hình đã nghiên cứu trước đó sẽ được đưa vào thực tế. Như đã đề cập trước đó, Streamlit giúp các nhà khoa học dữ liệu triển khai ứng dụng web cũng như thực hiện các công việc trực quan hóa dữ liệu một cách nhanh chóng. Điểm

đặc biệt của Streamlit được thể hiện ở chỗ là nó có thể triển khai chỉ với ít dòng code mà không cần tập trung quá nhiều vào việc cần phải có kỹ thuật cũng như kiến thức về lập trình front-end. Đề tài sẽ cần đến sự hỗ trợ thư viện Streamlit của Python.

Việc cần ở cho bài toán là phải xây dựng hàm tiền xử lý và trích xuất đặc trưng cho ứng dụng. Đồng thời, nghiên cứu cũng cần xuất ra hai mô hình tương ứng với hai đầu ra đã được huấn luyện trước đó để dự đoán cho dữ liệu mới được nhập vào. Sau đó, việc thiết kế dashboard để trực quan hóa dữ liệu sẽ được thực hiện. Cuối cùng kết quả mà đề tài thu được ở đây là một hình ảnh một bộ dữ liệu đã được dự đoán và dashboard trực quan về phản ứng của khách hàng đối với sản phẩm thông qua các đánh giá. Quá trình hoạt động của ứng dụng sẽ được thực hiện qua một loạt các bước như sau. Ban đầu, người dùng nhập một liên kết về sản phẩm mà người dùng muốn ứng dụng phân tích. Sau đó, ứng dụng sẽ tự động thu thập dữ liệu từ liên kết của chính sản phẩm đó. Dữ liệu thu được sẽ được hiển thị trên ứng dụng và là dữ liệu thô. Sau đó chúng sẽ được tiền xử lý, trích xuất đặc trưng, và được áp dụng mô hình để phân tích. Cuối cùng, nghiên cứu sẽ thu được hình ảnh trực quan hóa thông qua một dashboard, giúp người dùng dễ dàng hiểu và tiếp nhận được xu hướng, phản ứng chung của bộ dữ liệu.

- Thiết kế hàm nhập


Analyzing sentiment from product reviews

Input your product link:

<https://hasaki.vn/san-pham/ca-phe-dak-lak-tay-da-chet-toan-than-cocoon-200ml-84643.html>

Analyze

Việc đầu tiên mà nghiên cứu cần làm ở đây là xây dựng hàm nhập liên kết của sản phẩm. Hàm nhập này sẽ tiếp nhận đầu vào là liên kết của sản phẩm, sau đó sẽ duyệt vào liên kết để đó để truy cập vào trang thông tin chứa dữ liệu đánh giá về sản phẩm. Dữ liệu đánh giá sản phẩm của khách hàng sẽ được thu thập một cách tự động, lặp qua các trang đánh giá dần dần cho đến kết thúc. Các thông tin cần thu thập cũng có cấu trúc giống như cấu trúc bộ dữ liệu đã thu thập trước đó, chẳng hạn như tên người đánh giá, nội dung, điểm số cho sản phẩm, và thời gian thực hiện việc đánh giá sản phẩm. Các thông tin trên đều là thông tin cần thiết cho việc phân tích cảm xúc.

 RUNNING... Stop Deploy

Analyzing sentiment from product reviews

Input your product link:

<https://hasaki.vn/san-pham/ca-phe-đak-lai-tay-da-chet-toan-than-cocoon-200ml-84643.html>

Analyze

Analysis Result

	name_comment	content_comment	predicted_sentiment	predicted_aspect	datetime_comment	product_variant
0	Anh Pham	sài ok. Mùi rất thích	positive	smell	21: 49 24/07/2020	Cà Phê Đắk Lắk Tây Da Chết Toàn Thân Cocoon 200ml đã mua hàng online
1	Châu Mỹ	thơm mùi cà phê, tẩy da chết tốt thơm mịn ở lưng dần dần nhạt, làm xong da mịn vô	negative	smell	21: 07 13/06/2020	Cà Phê Đắk Lắk Tây Da Chết Toàn Thân Cocoon 200ml đã mua hàng online
2	Dương Nguyễn Mina	tuyệt vời	positive	stone	16: 39 01/06/2020	Cà Phê Đắk Lắk Tây Da Chết Toàn Thân Cocoon 200ml đã mua hàng online
3	Khánh Quỳnh	Đội cả nhân mình thì đây là loại tẩy tế bào chết cho body đáng để sử dụng. Chất của t	positive	smoothing	20: 50 19/05/2020	Cà Phê Đắk Lắk Tây Da Chết Toàn Thân Cocoon 200ml đã mua hàng online
4	Nguyễn Danh Thái		neutral	others	19: 59 07/04/2020	Cà Phê Đắk Lắk Tây Da Chết Toàn Thân Cocoon 200ml đã mua hàng online
5	Tuệ Nhi	xả cực tốt lun	positive	others	19: 06 19/03/2020	Cà Phê Đắk Lắk Tây Da Chết Toàn Thân Cocoon 200ml đã mua hàng online
6	Linh Phương	Bảng để mua lại	positive	others	01: 18 23/02/2020	Cà Phê Đắk Lắk Tây Da Chết Toàn Thân Cocoon 200ml đã mua hàng online
7	TRÚC	Highly recommend! Nên mua, dùng hàng tuần ít nhất 1 lần để tẩy da chết cho lưng, c	positive	texture	14: 45 18/01/2021	Cà Phê Đắk Lắk Cocoon Lầm Sạch Da Chết Toàn Thân 200ml đã mua hàng onli
8	Mỹ Ngọc	rất thơm luôn, hạt cà phê thực ra cũng ko lớn tới mức làm thương tổn da, sau khi xả d	neutral	smoothing	20: 28 11/01/2021	Cà Phê Đắk Lắk Cocoon Lầm Sạch Da Chết Toàn Thân 200ml
9	Hạnh Nguyễn		neutral	others	13: 08 06/01/2021	Cà Phê Đắk Lắk Cocoon Lầm Sạch Da Chết Toàn Thân 200ml đã mua hàng onli

Hình 25: Kết quả của Website

Cuối cùng, sau khi thực hiện việc nhập liên kết, thì dữ liệu từ quá trình thu thập cũng được hiển thị dưới dạng bảng và có cấu trúc tương đương với cấu trúc bảng của dữ liệu đã thu thập từ đầu bài toán.

KẾT LUẬN

Tóm lại, nghiên cứu đã mang lại những kết quả và hướng phát triển rộng mở phục vụ cho công việc nghiên cứu sau này. Qua quá trình thực hiện, đề tài đã nhận thấy rằng việc áp dụng thuật toán để phân tích cảm xúc có thể đem lại kết quả đáng tin cậy. Mô hình này có tính đơn giản, giúp triển khai bài toán nhanh chóng. Ngoài ra, việc sử dụng thư viện Streamlit đã giúp bài toán trực quan hóa dữ liệu đầu ra với một giao diện thân thiện, dễ sử dụng cho người dùng. Với bảng dashboard từ giai đoạn trực quan hóa, người dùng có thể phân tích các xu hướng phản ứng, đánh giá đối với sản phẩm. Đối với các nhà quản trị doanh nghiệp, điều này giúp doanh nghiệp phát hiện các vấn đề nảy sinh để đưa ra phương án giải quyết một cách kịp thời. Điều này thực sự ảnh hưởng đến sự sống còn của sản phẩm, và sâu xa hơn là sự tồn tại doanh nghiệp.

Những hạn chế của báo cáo:

- Quy trình xử lý dữ liệu văn bản chưa được hoàn chỉnh. Do các bình luận trên các nền tảng hiện nay có nhiều kiểu viết như viết tắt, teencode,... và các lỗi sai chính tả trong lúc gõ chữ như viết thiếu dấu, viết dư ký tự,... Gây khó khăn cho việc xử lý ngôn ngữ tự nhiên.
- Việc gán nhãn dữ liệu chưa chính xác hoàn toàn, còn nhiều nhãn được gán sai.
- Bài báo cáo trên sử dụng các loại dấu câu để phân tách câu, phương pháp này chưa hợp lý.

Hướng phát triển tương lai của đề tài:

- Cải thiện việc xử lý ngôn ngữ tự nhiên.

- Nâng cao độ chính xác trong gắn nhãn dữ liệu.
- Xây dựng mô hình phân loại bình luận bằng các thuật toán phức tạp và có độ chính xác cao hơn.
- Phân tách câu theo ngữ nghĩa của câu.
- Sử dụng các quy trình thu thập dữ liệu, xử lý dữ liệu, mô hình phân loại bình luận khách hàng đã xây dựng ở trên để tạo dashboard tự động hóa.

TÀI LIỆU THAM KHẢO

1. Chollet, F. (2015). *Keras: The Python deep learning library*.
<https://keras.io/>
2. Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), 51–89.
<https://doi.org/10.1002/aris.1440370103>
3. Google AI for Developers. (n.d.). *About generative models*.
<https://ai.google.dev/gemini/gemini-api/docs/overview>
4. Google AI for Developers. (n.d.). *Gemini API overview*.
<https://ai.google.dev>
5. Hasaki.vn. (n.d.). *Trang chính thức của Hasaki Beauty & Clinic*.
<https://hasaki.vn>
6. Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/>
7. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.
<https://arxiv.org/abs/1412.6980>
8. Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.
<https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>
9. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, arXiv:1301.3781. <https://arxiv.org/abs/1301.3781>
10. Nápoles-Duarte, A. M., Castillo-Barrera, F. E., & Tovilla-Hernández, C. (2022). Streamlit: Democratizing machine learning

app development for everyone. In *Proceedings of the LatinX in AI Research at ICML 2022*. <https://arxiv.org/abs/2211.13339>

11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

<http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

12. Selenium. (n.d.). *Selenium WebDriver documentation*.

<https://www.selenium.dev/documentation/>

13. TopDev. (2023). *Data mining là gì?*. <https://topdev.vn/blog/data-mining-la-gi/>

14. What Is Sentiment Analysis? (2020). *MonkeyLearn Blog*.

<https://monkeylearn.com/sentiment-analysis/>